TABLA DE CONTENIDOS

TABLA DE ILUSTRACIONES

1 INSTALACIÓN Y PUESTA EN MARCHA

Los requisitos mínimos, tanto hardware como software, para poder ejecutar la aplicación son los siguientes:

Requisitos Hardware	
CPU	400 MHz o superior (recomendable 1.5GHz). El tipo de procesador es indiferente. Lo único necesario es que exista una distribución de la Máquina Virtual Java para dicho procesador.
Memoria	512 MB (recomendable 1GB).
Disco duro	3 MB libres (sin JRE).

TABLA 1 REQUISITOS HARDWARE

Requerimientos Software	
Sistema operativo	Cualquier sistema operativo para el que se exista una Máquina Virtual Java.
Máquina Virtual Java	Java SE 6.0 o superior. (http://java.sun.com).

TABLA 2 REQUISITOS SOFTWARE

Los pasos para la instalación y puesta en marcha se detallan a continuación:

1) Copiar el contenido de la carpeta `/BioClass` a una carpeta en el disco duro. El contenido de esa carpeta es el siguiente:

   - `conf`. Contiene los archivos de configuración de AIBench.
   - `lib`. Contiene todas las librerías usadas para el desarrollo del proyecto.
   - `plugins_bin`. Contiene los archivos .jar del proyecto.
   - `plugins_src`. Contiene el código fuente del proyecto.
   - Resto de ficheros para su correcto funcionamiento, entre los cuales se encuentran lanzadores para los distintos sistemas operativos

2) Para la puesta en marcha se debe ejecutar el archivo *run.bat*, si el usuario se encuentra en un sistema operativo tipo Windows o *run.sh* en caso de Unix .

Una vez el usuario lanza la aplicación se mostrará una imagen de precarga. Durante este proceso se verifica el buen funcionamiento de la aplicación y el estado de los plugins actuales. Dado que AIBench es un Framework basado en plugins, uno de ellos es BioClass.

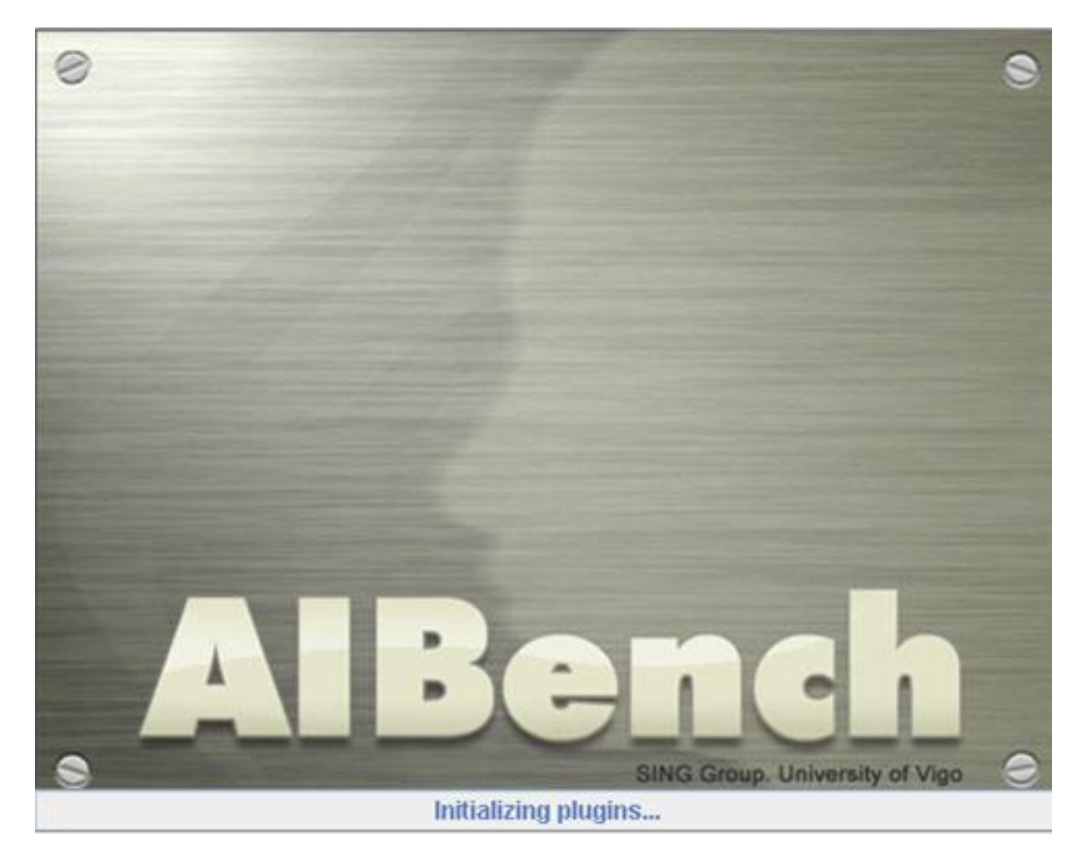


ILUSTRACIÓN 1 PANTALLA DE CARGA

Una vez finaliza la carga de módulos se mostrará la ventana principal de la aplicación.

1.1 ASPECTOS GENERALES DE LA INTERFAZ

El Framework de AIBench permite configurar la disposición de las ventanas de acción en función de los deseos de desarrollador. Para este caso particular, se ha reducido el número de las mismas, mostrando una visión más clara de los datos a mostrar.

Como se puede ver en la Ilustración 2, la interfaz está formada por la **Barra de Menús, el Panel del Clipboard, el Panel de log y el Panel de resultados.**

Veamos más detenidamente cada uno de los componentes de la Interfaz principal:

- **La Barra de menús** permite acceder a cada una de las operaciones disponibles en Bioclass: cargar datos, filtrarlos o clasificarlos en función de los deseos del usuario.

- **El Panel del Clipboard** contiene los elementos generados o utilizados por las operaciones, es decir, representa los datos a procesar por cada uno de los algoritmos. También sirve como medio de lanzamiento de las vistas asociadas a datos.

- **El Panel de resultados** muestra las vistas asociadas a un determinado tipo de dato. El proceso de Visualización es sencillo, basta con seleccionar el elemento que se desee visualizar en el *Arbol del Clipboard*.

- **El panel de logs** permite al usuario tener conocimiento de los procesos llevados a cabo y va mostrando los mensajes relativos a las operaciones que se están realizando.

Además de los paneles principales, existen otros dos adicionales anexados al de log: el *monitor de memoria*, el cual muestra un gráfico con los niveles de memoria ocupada y el panel *AIBench Shell*, permitiendo al usuario cargar, guardar y borrar flujos o experimentos.



ILUSTRACIÓN 2 INTERFAZ GENERAL

2 OPERACIONES DISPONIBLES EN BIOCLASS

Las operaciones en BioClass están agrupadas por categorías: *Corpus, Clasificación y Filtrado.*

Las operaciones disponibles bajo el *Menú Corpus* permiten cargar los conjuntos de datos que serán procesados por la aplicación. A través de *Clasificación* se pueden crear modelos de razonamiento, entrenarlos o testearlos de diferentes maneras. Por último, el *Menú de Filtrado* contiene aquellos algoritmos que permitan ajustar los conjuntos de datos en función de sus dimensiones.

2.1 GRUPO CORPUS

Como se ha comentado, el grupo corpus permite la carga de conjuntos de datos a partir de diferentes fuentes. En este caso concreto, se ha considerado solamente la carga de datos a partir de ficheros de texto plano. Es por ello que sólo se ha incluido una operación, *Cargar corpus desde fichero (Load corpus from file).* La siguiente imagen muestra una instantánea de dicho menú.

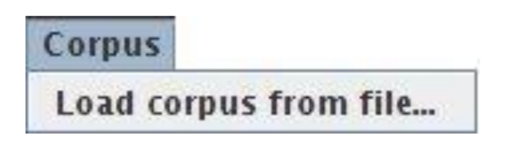


ILUSTRACIÓN 3 MENÚ GRUPO CORPUS

2.1.1 CARGAR CORPUS A PARTIR DE FICHERO

El conjunto de datos representa los documentos que se desean procesar a través de la aplicación BioClass. Soporta matrices de dispersión de datos como medio representativo, en formato *CSV* o *Arff*, siendo este último formato propio de la API de Weka.

Este proceso de operación se lleva a cabo mediante la opción de menú **Load corpus from file.** Una vez seleccionada se muestra la siguiente interfaz.

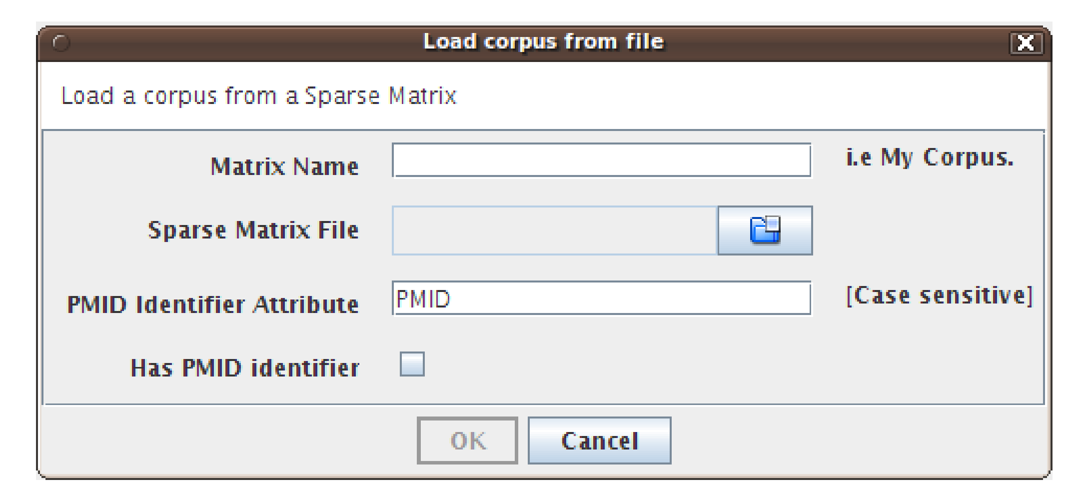


ILUSTRACIÓN 4 INTERFAZ CARGAR CORPUS

Como se puede observar, la ventana de operación presenta diferentes opciones de configuración. Veamos con más detenimiento cada una de ellas.

- **Matrix Name:** Representa el nombre mediante el cual se visualizará el *Corpus* en el árbol Clipboard, sirviendo como identificador del mismo ante operaciones y visores.

- **Sparse Matrix File:** Este parámetro contiene la ruta al conjunto de datos que se desea procesar. Actualmente están soportados dos tipos de formato CSV y Arff. Tal como se ha mencionado, el sistema es capaz de distinguir el tipo de fichero por el contenido del mismo.

- **PMID Identifier Attribute:** El corpus de datos, al estar formado por atributos (Palabras) e instancias (Documentos), en algunos casos, se necesita identificar cada uno de los documentos a través de un atributo clave. En el caso de los documentos Medline se utiliza el PMID. Este campo recoge el nombre de dicho identificador.

- **Has PMID Identifier:** Este parámetro es sumamente importante si se desea especificar un identificador de documento. En caso de haberse seleccionado, el algoritmo de carga tiene en cuenta el contenido del parámetro **PMID Identifier Attribute** y lo utiliza como clave. En caso de que el atributo no se encuentre contenido en el conjunto de datos, se notifica al usuario de la no existencia a través una ventana informativa.

2.2 Grupo Clasificación

A través del grupo de clasificación el usuario puede realizar tareas relacionadas con la creación de modelos de razonamiento, entrenamiento o testeo. BioClass soporta el diferentes clasificaciones: K vecinos más próximos, Naive Bayes o dos implementaciones diferentes de Maquinas de Soporte Vectorial (SVM).

Las opciones principales disponibles bajo el Menú de clasificación se muestran a continuación.

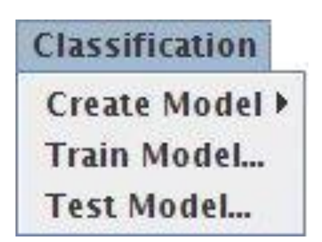


ILUSTRACIÓN 5 MENÚ CLASIFICACIÓN PRINCIPAL

A parte de las opciones *Entrenar Modelo (Train Model)* y *Testear Modelo (Testear Modelo)*, la opción *Crear modelo (Create Model)* ofrece la posibilidad de desplegarse hacia la derecha, dando paso a las opciones de creación de modelos de razonamiento.

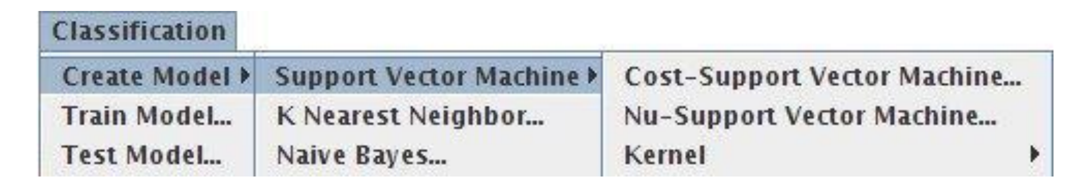


ILUSTRACIÓN 6 MENÚ CREACIÓN DE MODELOS DE RAZONAMIENTO

2.2.1 CREAR MODELO K NEAREST NEIGHBOR (KNN)

La operación de *K Nearest Neighbor* permite al usuario crear un clasificador del tipo supervisado. Este tipo de clasificador se basa en el cálculo de distancias entre documentos para la posterior clasificación en base a su vecindad con el conjunto de entrenamiento.

Una vez seleccionada la opción desde el Menú, se la lanzará la siguiente ventana.

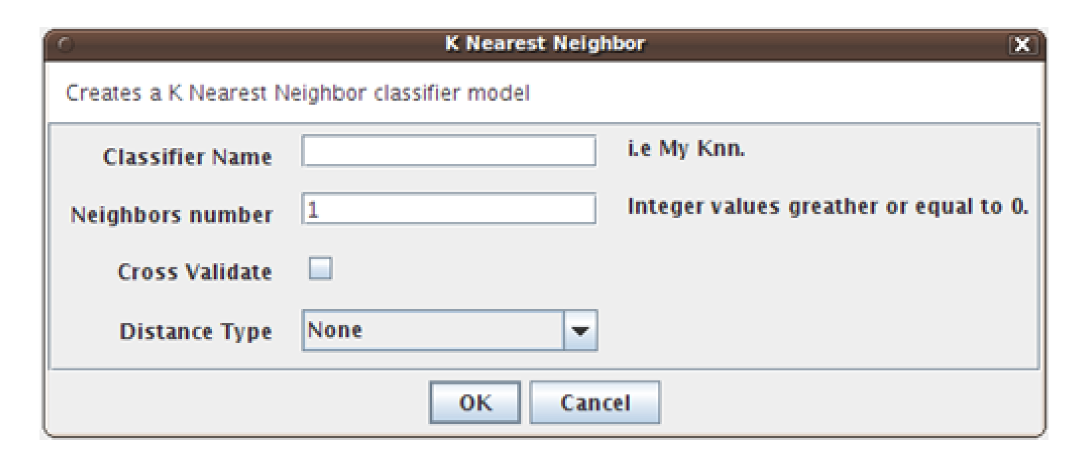


ILUSTRACIÓN 7 CREAR MODELO KNN

Los parámetros de los que dispone la operación son los siguientes:

- **Classifier Name:** Se corresponde con el identificador del elemento en el *Panel del Clipboard*, como también para las operaciones.

- **Neighbors number:** Se corresponde con el número de vecinos a utilizar por el clasificador. Este parámetro afecta directamente al número de vecinos que se utilizará durante el proceso votación y asignación de una instancia a una determinada clase. Los valores permitidos son enteros mayores que 0, en caso de introducir valores incorrectos se notifica al usuario a través de la correspondiente ventana de error.

- **Cross Validate:** Este parámetro permite al modelo, clasificar las instancias en función del algoritmo *Hold One Out* (Ver sección 3.1 Soluciones técnicas adoptadas del manual técnico). De esta manera se selecciona al mejor *K* valor de entre todos ellos en vez de realizar una votación entre los vecinos.

- **Distance Type:** El desplegable permite seleccionar el algoritmo del cálculo de distancias entre los vecinos. Las opciones disponibles son: Ninguna, Inversa de la distancia y 1-Distancia (Reversa)

2.2.2 CREAR MODELO NAIVE BAYES

La operación de *Naive Bayes* permite al usuario crear un clasificador probabilístico basado en el teorema de *Bayes*. Una vez se selecciona la opción del Menú, se lanza la siguiente ventana.

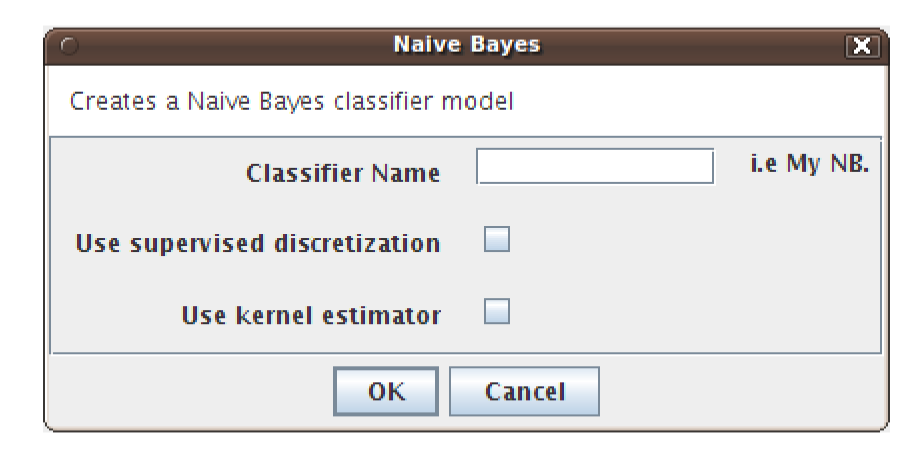


ILUSTRACIÓN 8 CREAR MODELO NAIVE BAYES

Como se puede ver en la imagen, la operación dispone de tres parámetros definidos. Veamos el significado con más detenimiento.

- **Classifier Name:** Se corresponde con el identificador que se le dará al elemento en el *Panel del Clipboard*, como también para las operaciones.

- **Use supervised discretization:** Si se elige esta opción, el clasificador discretiza los valores del conjunto de datos para obtener valores nominales en vez de numéricos.

- **Use kernel estimator:** Si el usuario selecciona esta opción, se utiliza un estimador basado en núcleo en vez de una distribución normal.

Cabe destacar que los parámetros "Use supervised discretization" y "Use kernel estimator" son mutuamente excluyentes, por tanto solo se puede seleccionar una de las dos opciones.

2.2.3 Crear modelo Cost Support Vector Machine

La operación de *Crear Cost Support Vector Machine(Cost-Support Vector Machine),* a la cual se puede acceder desde el Submenú de *Support Vector Machine,* que a su vez se encuentra recogido dentro del grupo de *Clasificación,* permite crear una Maquina de soporte vectorial con parámetro de Coste. Una vez seleccionada la opción del menú, se lanza la siguiente interfaz.

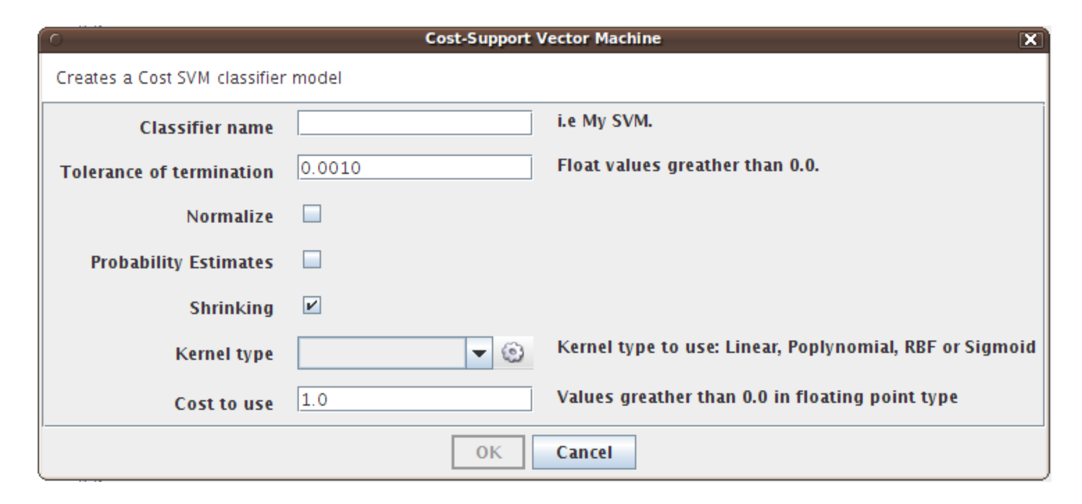


ILUSTRACIÓN 9 CREAR MODELO COST-SVM

Como se puede ver en la imagen, la operación dispone de múltiples parámetros configurables. Veamos el significado con más detenimiento.

- **Classifier Name:** Se corresponde con el identificador del elemento en el *Panel del Clipboard*, como también para las operaciones.

- **Tolerance of termination**: Este parámetro permite establecer un valor como criterio de terminación del algoritmo. Los valores permitidos son flotantes mayores que *0.0*. en caso de introducir valores incorrectos se notifica al usuario.

- **Normalize**: En caso de que el usuario decida utilizar este parámetro, el clasificador normalizará los datos correspondientes a los vectores.

- **Probability Estimates**: El proceso de estimación de una máquina de soporte vectorial, por defecto atribuye valores -1 o 1 a cada una de las instancias en función de la pertenencia a una clase o a otra. Si la estimación probabilística es seleccionada, se calculan probabilidades de pertenencia de cada una de las instancias respecto de las clases, en vez de atribuir valores absolutos.

- **Shrinking:** Este parámetro permite al clasificador utilizar la heurística de *Shrinking* en el proceso de clasificación. Por defecto se encuentra activado.

- **Kernel type:** usando este desplegable, el usuario puede seleccionar el Kernel a usar con la Máquina de Soporte Vectorial. El Kernel puede crearse *(Submenú de Kernels)* en un paso anterior, o mediante el menú contextual a la derecha del desplegable.

- **Cost**: El parámetro de coste permite al usuario ajustar los márgenes del *hiperplano* de la SVM, afinando la pertenencia de cada una de instancias en relación a las clases. Los valores permitidos son flotantes mayores que *0.0*, en caso de introducir valores incorrectos se no notificara se notificará al usuario.

2.2.4 Crear modelo Nu Support Vector Machine

La operación de *Crear Nu Support Vector Machine(Nu-Support Vector Machine),* a la cual se puede acceder desde el Submenú de *Support Vector Machine,* que a su vez se encuentra recogido dentro del grupo de *Clasificación,* permite crear una Máquina de Soporte Vectorial con parámetro *"Nu"* para el control del margen del *hiperplano*. Una vez, seleccionada la opción del menú, se lanza la siguiente interfaz.

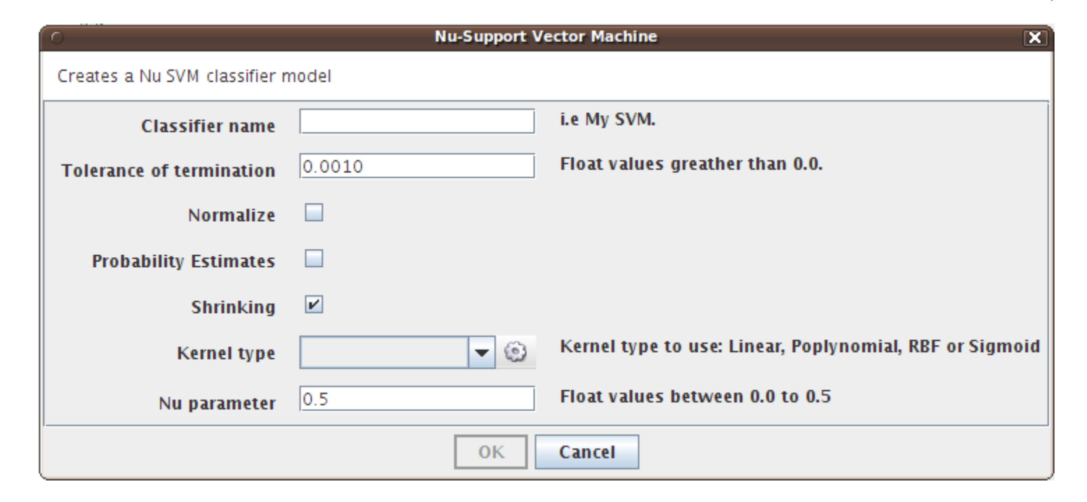


ILUSTRACIÓN 10 CREAR MODELO NU-SVM

Como se puede ver en la imagen, la operación dispone de múltiples parámetros configurables. Veamos el significado con más detenimiento.

- **Classifier Name:** Se corresponde con el identificador del elemento en el *Panel del Clipboard*, como también para las operaciones.

- **Tolerance of termination**: Este parámetro permite establecer un valor como criterio de terminación del algoritmo. Los valores permitidos son flotantes mayores que *0.0*. En caso de introducir valores incorrectos se notifica al usuario.

- **Normalize**: En caso de que el usuario decida utilizar este parámetro, el clasificador normalizará los datos correspondientes a los vectores.

- **Probability Estimates**: El proceso de estimación de una máquina de soporte vectorial, por defecto atribuye valores -1 o 1 a cada una de las instancias en función de la pertenencia a una clase o a otra. Si la estimación probabilística es seleccionada, se calculan probabilidades de pertenencia de cada una de las instancias respecto de las clases, en vez de atribuir valores absolutos.

- **Shrinking:** Este parámetro permite al clasificador utilizar la heurística de *Shrinking* en el proceso de clasificación. Por defecto esta activado.

- **Kernel type:** usando este desplegable, el usuario puede seleccionar el Kernel a usar con la Máquina de Soporte Vectorial. El Kernel puede crearse *(Submenú de Kernels)* en un paso anterior, o mediante el menú contextual a la derecha del desplegable

- **Nu**: Este parámetro, permite ajustar el límite superior del margen de error, además de comportarse también como límite inferior de la fracción de vectores. Los valores permitidos están comprendidos entre *0.5 y 0.0*. En caso de introducir valores incorrectos se notifica al usuario.

2.2.5 Kernels para Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial utilizan funciones Kernel para trasladar los vectores representativos de cada instancia al espacio de características. La calidad de la SVM vendrá caracterizada por la parametrización y tipo de kernel, pues influirá de directamente los modelos de resolución lineal a aplicar.

BioClass soporta varios tipos de *Kernel*: *Lineal*, *Polinomial*, *Radial* o *Sigmoidal*. El usuario debe tener en cuenta cada uno de los parámetros de los que disponen los Kernels para afinar al máximo la SVM.

El acceso al *Submenú de Kernels* se pude observar a través de la siguiente imagen.

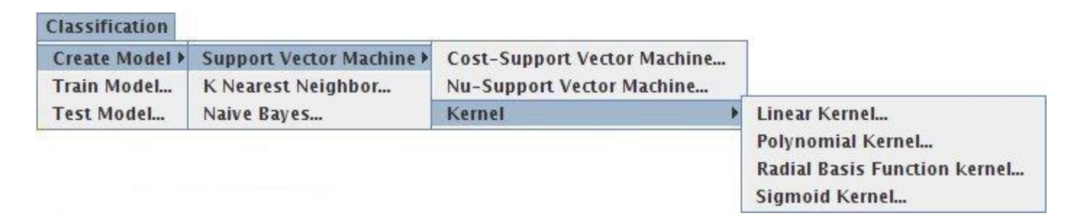


ILUSTRACIÓN 11 SUBMENU KERNELS

2.2.5.1 Crear kernel Lineal

Esta operación permite crear un Kernel Lineal para una Support Vector Machine basándose en la siguiente ecuación:

$$Kernel(u, v) = \langle u, v \rangle$$

Los parámetros *u* y *v* representaran cada una de las instancias pertenecientes al *Corpus*. Una vez seleccionada la opción desde del *Submenú*, se lanza la siguiente ventana.

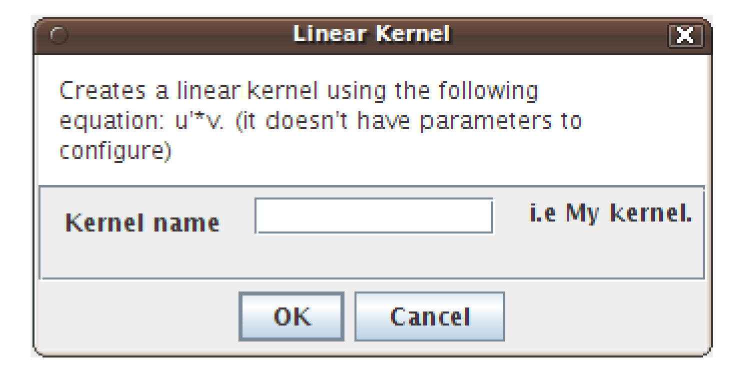


ILUSTRACIÓN 12 CREAR KERNEL LINEAL

Como se puede ver en este caso, al intervenir solamente los vectores de las instancias, no posee ningún parámetro adicional. Por tanto, sólo se tiene en cuenta el nombre del elemento.

- **Kernel name:** Se corresponde con el identificador del elemento en el *Panel del Clipboard*, como también para las operaciones.

2.2.5.2 CREAR KERNEL POLINOMIAL

Esta operación permite crear un Kernel Polinomial para una Support Vector Machine. Este se basa en la siguiente ecuación:

$$Kernel(u, v) = (gamma \cdot u' \cdot v + coef)^{degree}$$

Los parámetros *u* y *v* representaran cada una de las instancias pertenecientes al *Corpus*. Veamos el resto de parámetros una vez seleccionada la operación desde el Submenú.

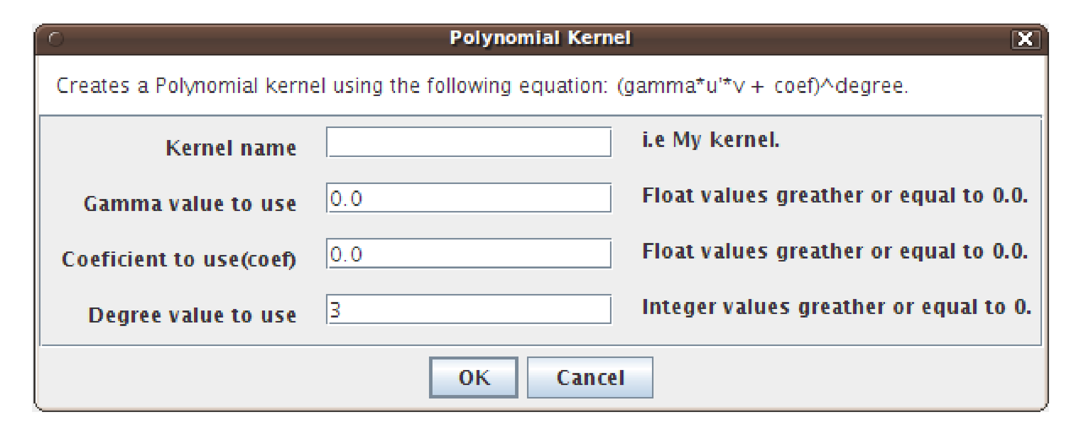


ILUSTRACIÓN 13 CREAR KERNEL POLINOMIAL

- **Kernel name:** Se corresponde con el identificador del elemento en el *Panel del Clipboard*, como también para las operaciones.

- **Gamma value to use:** El valor de este parámetro sustituirá al *Gamma* existente en la ecuación del Kernel. Solamente se permiten valores flotantes mayores o iguales a *0.0.* En caso de no supeditarse a las restricciones, se mostrará una ventana informativa.

- **Coeficient to use:** El valor de este parámetro representa a la variable *Coef* perteneciente a la ecuación del Kernel. Solamente se permiten valores flotantes mayores o iguales a *0.0.* En caso de no supeditarse a las restricciones, se mostrará una ventana informativa.

- **Degree value to use:** Valor del parámetro *Degree* que será sustituido en la ecuación del Kernel. Solamente se permiten valores enteros mayores o iguales a *0.* En caso de no supeditarse a las restricciones, se mostrará una ventana informativa.

2.2.5.3 CREAR KERNEL CON BASE RADIAL

Esta operación permite crear un Kernel de base radial o RBF para una Support Vector Machine. Este se basa en la siguiente ecuación:

$$Kernel(u, v) = e^{-gamma \cdot |u \cdot v|^2}$$

Los parámetros *u* y *v* representan cada una de las instancias pertenecientes al *Corpus*. Veamos el resto de parámetros una vez seleccionada la operación desde el Submenú.

Radial Basis Function kernel
Creates a Radial Basis Function kernel using the following equation: exp(-gamma*|u-v|^2).
Kernel name | i.e My kernel.
Gamma value to use 0 | Float values greather or equal to 0.0.
OK Cancel

ILUSTRACIÓN 14 CREAR KERNEL RADIAL

- **Kernel name:** Se corresponde con el identificador del elemento en el *Panel del Clipboard*, como también para las operaciones.

- **Gamma value to use:** El valor de este parámetro sustituirá al *Gamma* existente en la ecuación del Kernel. Solamente se permiten valores flotantes mayores o iguales a *0.0.* En caso de no supeditarse a las restricciones, se mostrará una ventana informativa.

2.2.5.4 CREAR KERNEL SIGMOIDAL

La operación permite crear un Kernel Sigmoidal para una Support Vector Machine. Este se basa en la siguiente ecuación:

$$Kernel(u, v) = \tanh(gamma \cdot u' \cdot v + coef)$$

Los parámetros *u* y *v* representan cada una de las instancias pertenecientes al *Corpus*. Veamos el resto de parámetros una vez seleccionada la operación desde el Submenú.

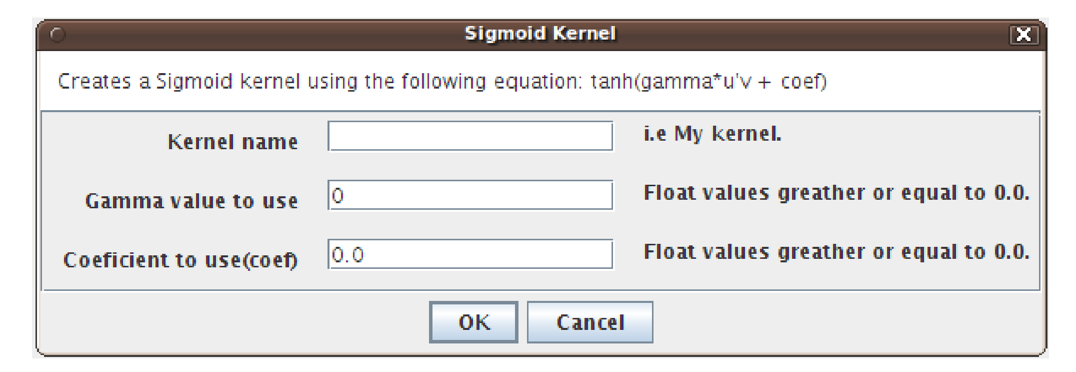


ILUSTRACIÓN 15 CREAR KERNEL SIGMOIDAL

- **Kernel name:** Se corresponde con el identificador del elemento en el *Panel del Clipboard*, como también para las operaciones.

- **Gamma value to use:** El valor de este parámetro sustituirá al *Gamma* existente en la ecuación del Kernel. Solamente se permiten valores flotantes mayores o iguales a *0.0.* En caso de no supeditarse a las restricciones, se mostrará una ventana informativa.

- **Coeficient to use:** El valor de este parámetro representa a la variable *Coef* perteneciente a la ecuación del Kernel. Solamente se permiten valores flotantes

mayores o iguales a *0.0.* En caso de no supeditarse a las restricciones, se mostrará una ventana informativa.

2.2.6 ENTRENAR MODELO

La operación de *Entrenar modelo (Train Model),* a la cual se puede acceder desde el grupo de *Clasificación,* permite entrenar un modelo de razonamiento, creado con anterioridad, sobre un conjunto de datos determinado. Este último también debe cargarse mediante la opción correspondiente en el *Menú Corpus*. Una vez, seleccionada la opción del menú, se lanza la siguiente interfaz.

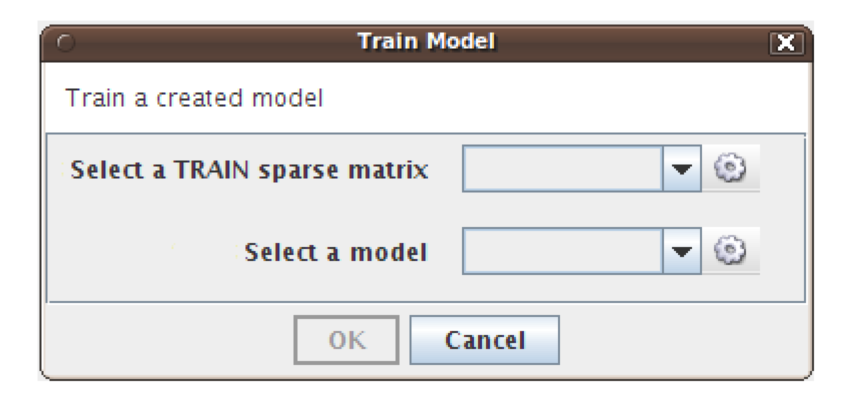


ILUSTRACIÓN 16 ENTRENAR UN MODELO

Como se puede ver en la imagen, la operación dispone de dos parámetros definidos. Veamos el significado con más detenimiento.

- **Select a TRAIN sparse matrix:** A través de este parámetro, el usuario puede seleccionar entre los distintos corpus cargados con anterioridad para utilizar como conjunto de entrenamiento. Cada uno de ellos se encuentra, como se ha comentado, en el *Panel del Clipboard*. En caso de no disponer de ningún conjunto de datos cargado previamente, se puede optar por hacerlo a través de las opciones del *Menú Corpus* o usando el acceso directo que se encuentra a la derecha del desplegable.

- **Select a model**: Al igual que en el parámetro anterior, este también se presenta como un desplegable, desde el cual se puede seleccionar un Modelo de razonamiento creado con anterioridad. Este modelo será entrenado en base al corpus seleccionado como parámetro en el paso anterior.

2.2.7 TESTEAR MODELO

La operación de testeo del modelo permite clasificar un corpus secundario en función del conjunto de entrenamiento sobre el cual se entrenó el modelo de razonamiento. Al igual que *Entrenar modelo (Train Model), Testear modelo (Test Model)* forma parte de las operaciones englobadas bajo el *Grupo de clasificación*. Por tanto se podrá acceder a ella a través de dicho Menú.

Una vez seleccionada la operación se presentará al usuario la siguiente ventana.

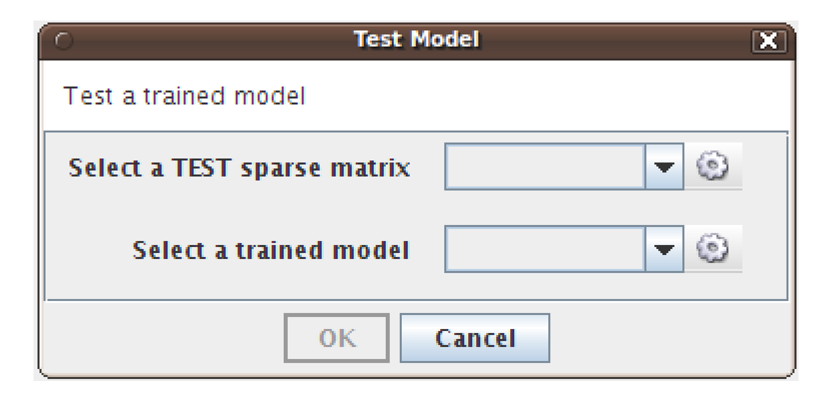


ILUSTRACIÓN 17 TESTEAR MODELO

Como se puede ver en la imagen, la operación dispone de dos parámetros definidos. Veamos el significado con más detenimiento.

- **Select a TEST sparse matrix:** Usando este desplegable, el usuario puede seleccionar entre los corpus que han sido cargados en la aplicación. Este corpus se utilizará como conjunto de datos a ser clasificado. Mencionar que se pude utilizar el menú contextual disponible a la derecha del desplegable para cargar nuevos Corpus.

- **Select a trained model**: Este parámetro permite seleccionar uno de los modelos anteriormente entrenados. Estos deben estar disponibles a través del *Panel del Clipboard*. De no ser así, se puede utilizar el menú contextual o la opción correspondiente de la *Barra de menús.*

2.3 GRUPO FILTRADO

El menú de filtrado permite al usuario realizar operaciones sobre la dimensionalidad de los datos. De esta manera el usuario puede redistribuir el número de instancias pertenecientes a cada una de las clases, como también comprimir el conjunto de atributos relevantes a cada corpus.

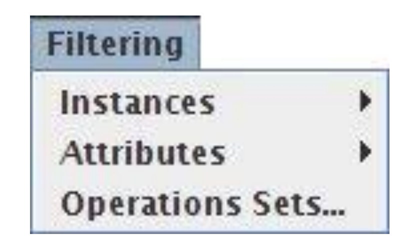


ILUSTRACIÓN 18 MENU FILTRADO

Como se puede ver en la imagen, los submenús de los que se dispone se corresponden directamente con la definición dada, pues el *Submenu Instances* permite redistribuir el número de instancias por clase, y el *Submenu Attributes* realiza operaciones similares pero a nivel de atributos. Por último, se ha añadido una *Operación basada en conjuntos (Operations Sets)* a nivel de atributos, para permitir el procesado de las matrices en caso de que su tamaño sea elevado. A continuación se muestran unas instantáneas de ambos submenús.

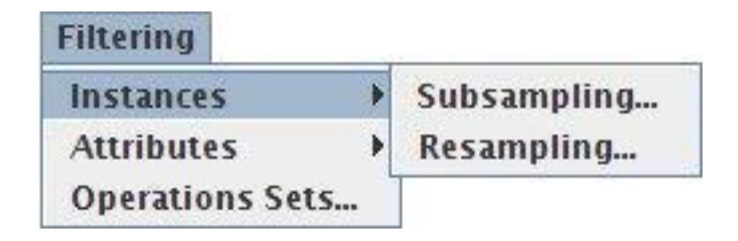


ILUSTRACIÓN 19 SUBMENÚ FILTADO DE INSTANCIAS

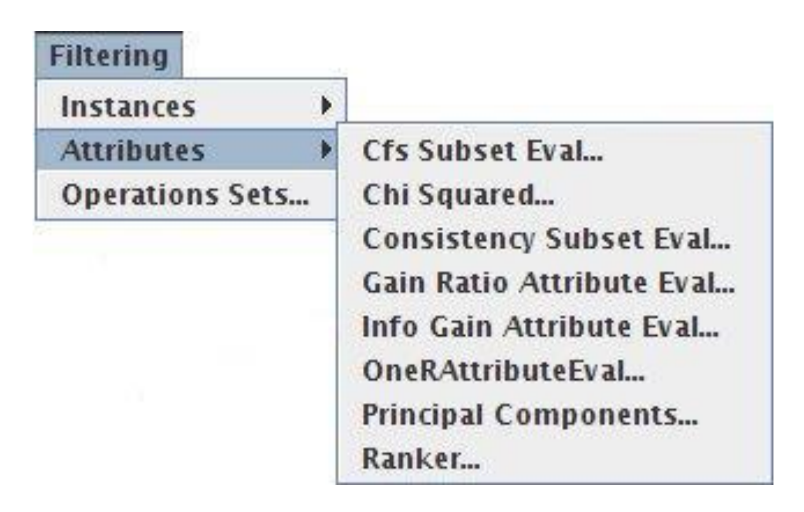


ILUSTRACIÓN 20 SUBMENÚ FILTRADO DE ATRIBUTOS

2.3.1 Operación de Subsampling

La operación de *Sub muestreo (Subsampling)* permite realizar una redistribución del número de instancias pertenecientes a cada una de las clases. El proceso podrá disminuir los documentos pertenecientes a la clase mayoritaria en función de la envergadura de la minoritaria. Este tipo de algoritmos sólo puede aplicarse a conjuntos de Corpus que vayan a servir de conjunto de entrenamiento para un clasificador.

La operación se encuentra accesible a través del *Submenú Instances,* dentro del grupo *Filtering.* A continuación se muestra la ventana de configuración.

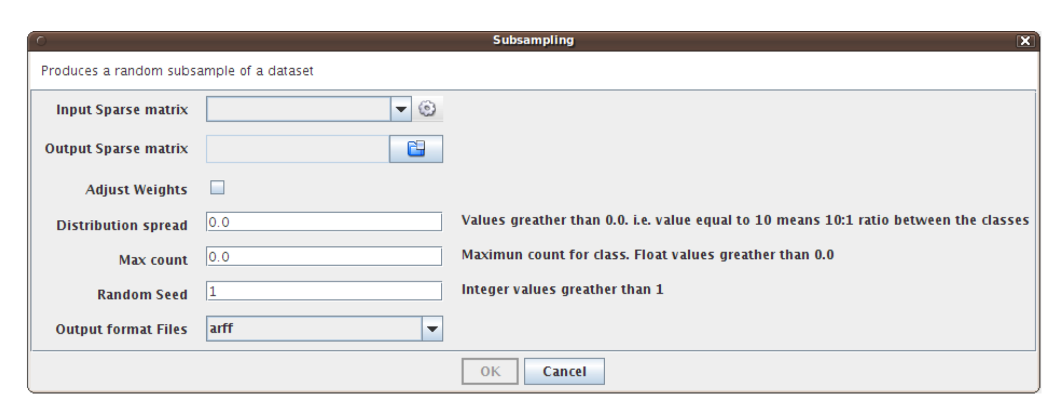


ILUSTRACIÓN 21 OPERACIÓN SUBSAMPLING

- **Input SparseMatrix:** usando el desplegable, el usuario podrá seleccionar entre los corpus que han sido cargados en la aplicación. Este corpus se utiliza como conjunto de datos a ser filtrado. Mencionar que se puede utilizar el menú contextual, disponible a la derecha del desplegable, para cargar nuevos Corpus.

- **Output Sparse Matrix:** Este parámetro representa el fichero de datos en el cual se escribirá el corpus una vez filtrado. Es decir, será el dato de salida de la operación.

- **Adjust Weights:** El ajuste de pesos permite al filtro redistribuir el peso global de las clases entre cada una de las instancias que la conforman. Por tanto se ha de tener especial cuidado en su utilizacion, ya que podría ocasionar un incremento del error.

- **Distribution spread**: A través de este parámetro se podrá establecer la distribución del número de instancias pertenecientes a cada una de las clases. Un valor igual a 10 equivaldría a establecer una distribución *10 a 1*, teniendo en cuenta que la clase mayoritaria seria *10* veces mayor a la otra. La restricción de valores queda supeditada a mayores o iguales a *0.0*. No obstante, si el usuario selecciona un valor igual a *0.0* el filtro no afectaría a los datos.

- **Max count:** Mediante este parámetro se especifica el número máximo de instancias por clase. La restricción de valores queda supeditada a mayores o iguales a *0.0*. Un valor igual a *0.0* implica ilimitado.

- **Random Seed:** Dado que el proceso de eliminación de instancias se realiza de forma pseudo-aleatoria, valores pertenecientes a este parámetro servirán como semilla para la generación de los números de instancias que se desecharán.

- **Output format files:** Permite elegir entre dos tipos de salida *CSV* y *Arff*. Por defecto el formato se ha preestablecido a *Arff.*

2.3.2 Operación de Resampling

La operación de *Re-muestreo (Resampling),* combina las técnicas de *Sub-muestreo* y *Sobre-muestro,* de modo que permite al usuario redistribuir las instancias pertenecientes a cada una de las clases hasta conseguir una distribución binomial entre ambas.

Veamos la interfaz de usuario que se presenta al seleccionar la operación. A esta se puede acceder a través del *Submenú Instances* y perteneciente al grupo de Filtrado.

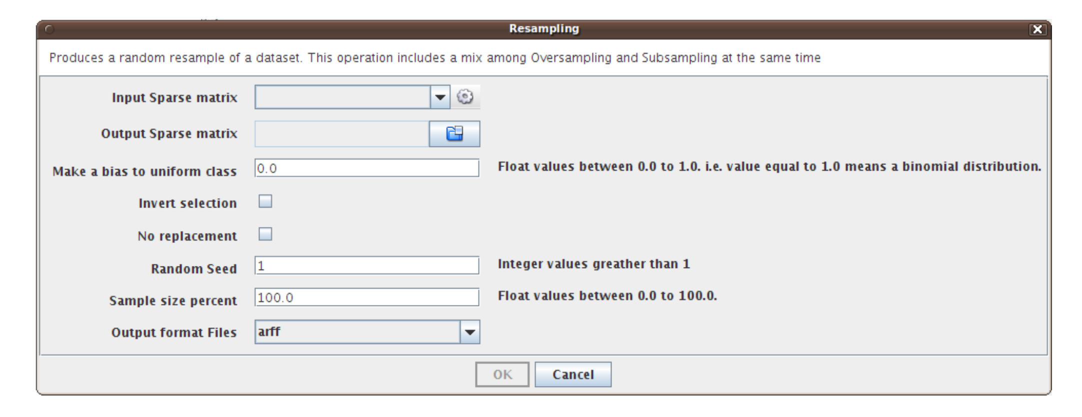


ILUSTRACIÓN 22 OPERACIÓN RESAMPLING

- **Input Sparse Matrix:** usando el desplegable, el usuario puede seleccionar entre los corpus que han sido cargados en la aplicación. Este corpus se utilizará como conjunto de datos a ser filtrado. Mencionar que se puede utilizar el menú contextual, disponible a la derecha del desplegable, para cargar nuevos Corpus.

- **Output Sparse Matrix:** Este parámetro representa el fichero de datos en el cual se escribirá el corpus una vez filtrado. Es decir será el dato de salida de la operación.

- **Make a bias to uniform class**: A través de este parámetro se puede establecer la distribución del número de instancias pertenecientes a cada una de las clases. Un valor igual a *1.0* equivaldría a establecer una distribución binomial entre las dos clases. La restricción de valores queda supeditada entre *0.0 y 1.0*. Si el usuario selecciona un valor igual a *0.0*, el filtro no afectaría a los datos.

- **Invert selection:** Si el usuario activa esta opción, el filtro invierte el proceso de selección que se utiliza como base para redistribuir el número de instancias de las clases. Este parámetro sólo es efectivo si no se produce reemplazo.

- **No replacement:** Este parámetro deshabilita el reemplazo de las instancias en el proceso de redistribución.

- **Random Seed:** Dado que el proceso de eliminación de instancias se realiza de forma pseudo-aleatoria, los valores pertenecientes a este parámetro servirán como semilla para la generación de los números de instancias que se desecharán.

- **Sample size percent:** Este parámetro permite al usuario establecer el porcentaje de la muestra sobre la que se trabajará. La restricción de valores queda supeditada entre *0.0 y 100.0.* El valor por defecto se encuentra establecido al 100% de la muestra.

- **Output format files:** Permite elegir entre dos tipos de salida *CSV* y *Arff*. Por defecto el formato se ha preestablecido a *Arff.*

2.3.3 Operación de conjuntos

La siguiente operación permite al usuario realizar un filtrado de atributos sobre un conjunto de datos, en función de la relación de existencia de los mismos. Utilizando la teoría de conjuntos y las relaciones de Intersección, diferencia y unión, se ha conseguido reducir la dimensionalidad de aquellos conjuntos extremadamente grandes e imposibles de procesar.

Esta operación se encuentra disponible a través *Menú de Filtrado*. A continuación se muestra una instantánea de las opciones disponibles.

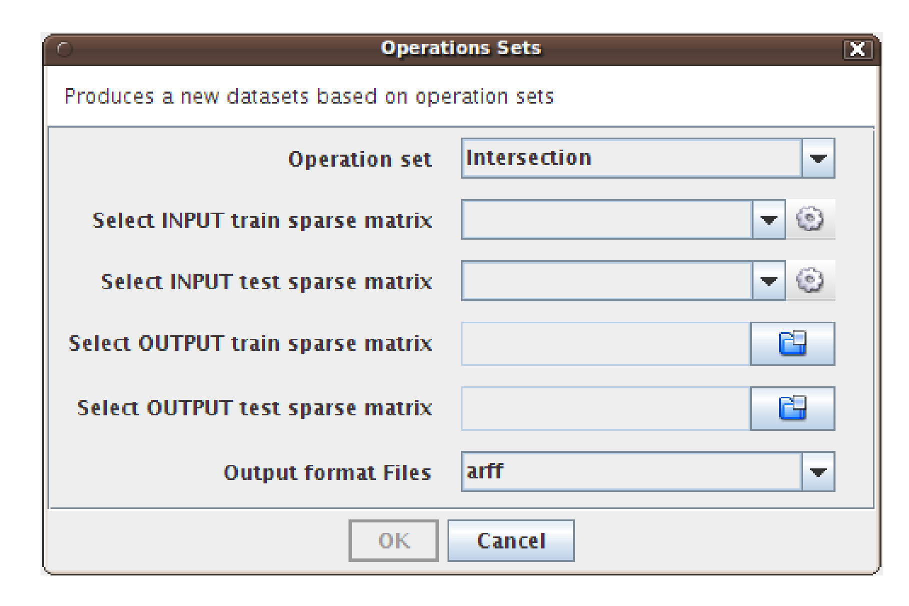


ILUSTRACIÓN 23 OPERACIÓN DE CONJUNTOS

- **Operación set:** El usuario puede seleccionar la operación de conjuntos a utilizar con los corpus. Las opciones disponibles son: *Intersección, Diferencia y Unión*. La opción por defecto está preestablecida a *Intersección*.

- **Select INPUT train sparse matrix:** Este parámetro se corresponde con el corpus que se utilizará como conjunto de entrenamiento. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select INPUT test sparse matrix:** Este parámetro se corresponde con el corpus que se utilizará como conjunto de testeo. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select OUTPUT train sparse matrix:** representará el fichero que se creará como nuevo corpus de entrenamiento una vez aplicado el filtro.

- **Select OUTPUT test sparse matrix:** representará el fichero que se creará como nuevo corpus de testeo una vez aplicado el filtro.

- **Output format files:** El usuario puede seleccionar mediante el desplegable el tipo de fichero de salida: CSV o Arff. Por defecto el formato se ha preestablecido a *Arff.*

2.3.4 Operación Cfs Subset Eval

La siguiente operación permite al usuario realizar un filtrado de atributos sobre un conjunto de datos haciendo uso del algoritmo *Cfs Subset Eval*. Los atributos seleccionados se evalúan teniendo en cuenta la capacidad predictiva de cada uno de ellos y su grado de redundancia común. Por consiguiente, se consigue un subconjunto final que poseerá un nivel expresivo similar al original.

Esta operación se encuentra disponible dentro del *Submenu Attributes*, el cual se encuentra contenido dentro *Menú de Filtrado*. A continuación se muestra una instantánea de las opciones disponibles.

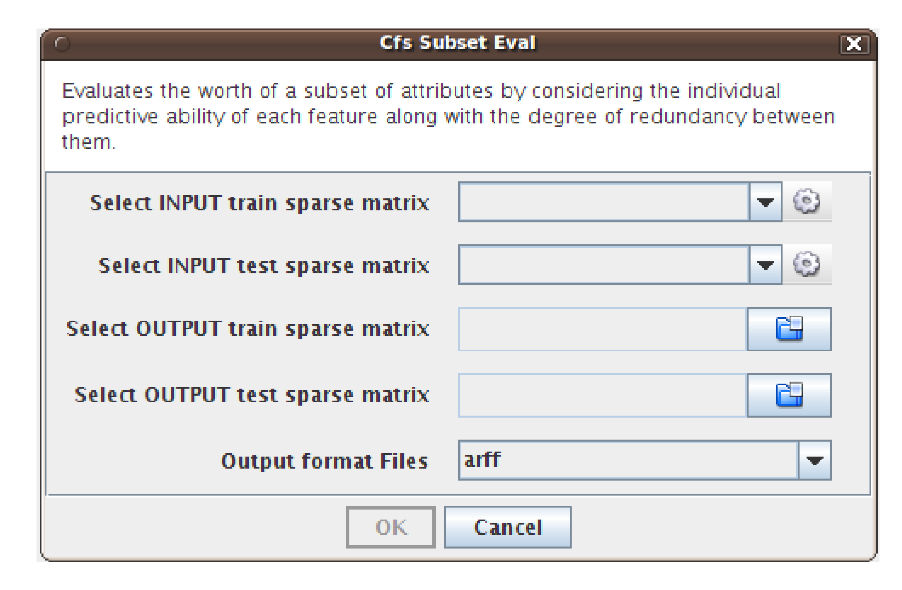


ILUSTRACIÓN 24 OPERACIÓN CFS SUBSET EVAL

- **Select INPUT train sparse matrix:** Este parámetro se corresponde con el corpus que se utilizará como conjunto de entrenamiento. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select INPUT test sparse matrix:** Este parámetro se corresponde con el corpus que se utilizará como conjunto de testeo. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select OUTPUT train sparse matrix:** Representa el fichero que se creará como nuevo corpus de entrenamiento una vez aplicado el filtro.

- **Select OUTPUT test sparse matrix:** Representa el fichero que se creará como nuevo corpus de testeo una vez aplicado el filtro.

- **Output format files:** El usuario puede seleccionar mediante el desplegable el tipo de fichero de salida: CSV o Arff. Por defecto el formato se ha preestablecido a *Arff.*

2.3.5 Operación Chi Squared

La siguiente operación permite al usuario realizar un filtrado de atributos sobre un conjunto de datos haciendo uso del algoritmo *Chi Squared*. Este evalúa cada atributo mediante el cálculo del valor estadístico *Chi-cuadrado* con respecto a la clase. Una vez obtenidos los valores de cada atributo se aplica un *Ranker*, reduciendo de este modo el conjunto inicial. Por consiguiente, se consigue un subconjunto final que poseerá un nivel expresivo similar al original.

Esta operación se encuentra disponible dentro del *Submenu Attributes*, el cual se encuentra contenido dentro *Menú de Filtrado*. A continuación se muestra una instantánea de las opciones disponibles.

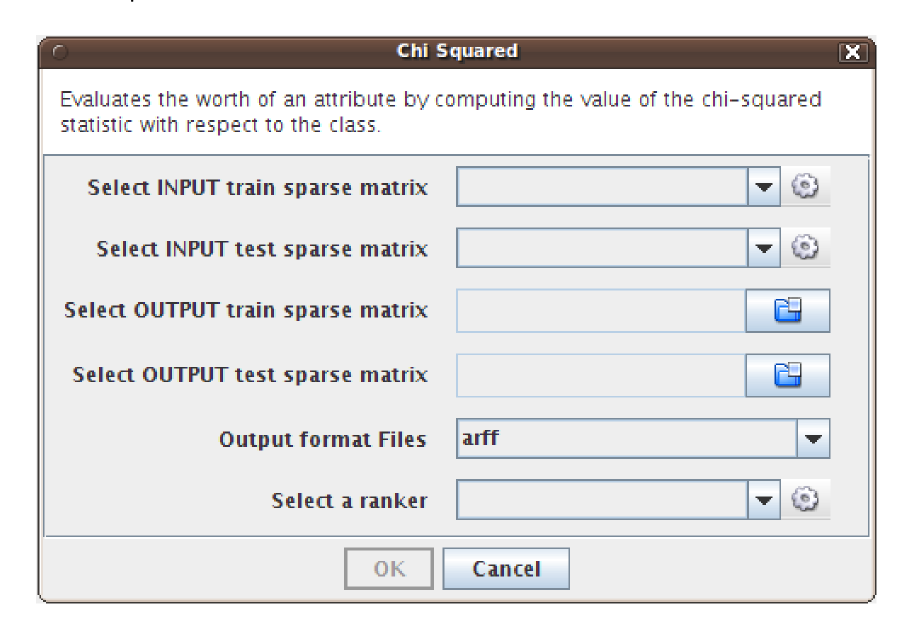


ILUSTRACIÓN 25 OPERACIÓN CHI SQUARED

- **Select INPUT train sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de entrenamiento. Usando el desplegable, el usuario podrá seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select INPUT test sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de testeo. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select OUTPUT train sparse matrix:** Representa el fichero que se creará como nuevo corpus de entrenamiento una vez aplicado el filtro.

- **Select OUTPUT test sparse matrix:** Representa el fichero que se creará como nuevo corpus de testeo una vez aplicado el filtro.

- **Output format files:** El usuario puede seleccionar mediante el desplegable el tipo de fichero de salida: CSV o Arff. Por defecto el formato se ha preestablecido a *Arff.*

- **Ranker:** Dado que el filtro necesita establecer un rango de valores aceptados por el usuario, a través de esta opción, el usuario puede seleccionar uno creado con anterioridad o servirse del menú contextual para generar uno nuevo.

2.3.6 Operación Consistency subset Eval

La siguiente operación permite al usuario realizar un filtrado de atributos sobre un conjunto de datos haciendo uso del algoritmo *Consistency Subset Eval*. Este evalúa el conjunto de atributos original en función de la coherencia de valores existentes en la clase, descartando aquellos que aporten poco valor a la misma. Por consiguiente, se consigue un subconjunto final que poseerá un nivel expresivo similar al original.

Esta operación se encuentra disponible dentro del *Submenu Attributes*, el cual se encuentra contenido dentro *Menú de Filtrado*. A continuación se muestra una instantánea de las opciones disponibles.

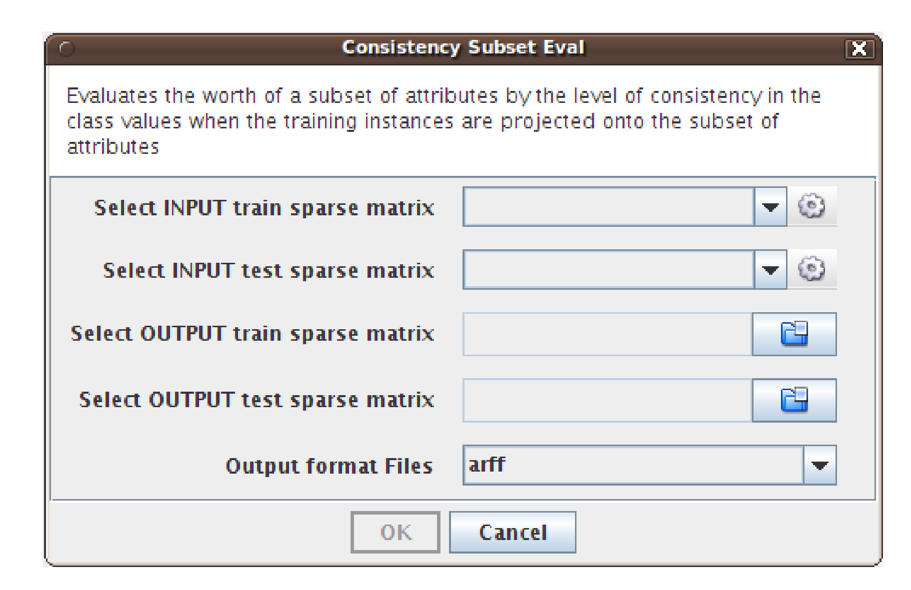


ILUSTRACIÓN 26 OPERACIÓN CONSISTENCY SUBSET EVAL

- **Select INPUT train sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de entrenamiento. Usando el desplegable, el usuario podrá seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select INPUT test sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de testeo. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select OUTPUT train sparse matrix:** representará el fichero que se creará como nuevo corpus de entrenamiento una vez aplicado el filtro.

- **Select OUTPUT test sparse matrix:** representará el fichero que se creará como nuevo corpus de testeo una vez aplicado el filtro.

- **Output format files:** El usuario puede seleccionar mediante el desplegable el tipo de fichero de salida: CSV o Arff. Por defecto el formato se ha preestablecido a *Arff.*

2.3.7 OPERACIÓN GAIN RATIO ATTRIBUTE EVAL

La siguiente operación permite al usuario realizar un filtrado de atributos sobre un conjunto de datos haciendo uso del algoritmo *Gain ratio attribute eval*. Este evalúa el peso de cada característica del conjunto de atributos original teniendo en cuenta la ganancia de información con respecto a la clase. Una vez obtenidos los valores de cada atributo se aplicará un *Ranker*, reduciendo el conjunto inicial. Por consiguiente, se consigue un subconjunto final que poseerá un nivel expresivo similar al original.

Esta operación se encuentra disponible dentro del *Submenu Attributes*, el cual se encuentra contenido dentro *Menú de Filtrado*. A continuación se muestra una instantánea de las opciones disponibles.

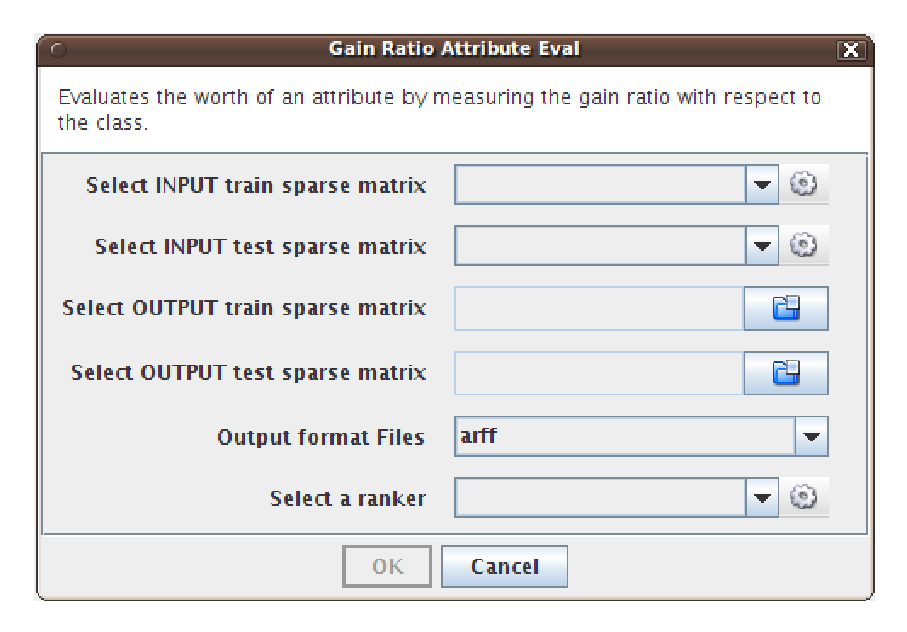


ILUSTRACIÓN 27 OPERACIÓN GAIN RATIO ATTRIBUTE EVAL

- **Select INPUT train sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de entrenamiento. Usando el desplegable, el usuario podrá seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select INPUT test sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de testeo. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select OUTPUT train sparse matrix:** Representa el fichero que se creará como nuevo corpus de entrenamiento una vez aplicado el filtro.

- **Select OUTPUT test sparse matrix:** Representa el fichero que se creará como nuevo corpus de testeo una vez aplicado el filtro.

- **Output format files:** El usuario puede seleccionar mediante el desplegable el tipo de fichero de salida: CSV o Arff. Por defecto el formato se ha preestablecido a *Arff.*

- **Ranker:** Dado que el filtro necesita establecer un rango de valores aceptados por el usuario, a través de esta opción, el usuario puede seleccionar uno creado con anterioridad o servirse del menú contextual para generar uno nuevo.

2.3.8 Operación One R attribute Eval

La siguiente operación permite al usuario realizar un filtrado de atributos sobre un conjunto de datos haciendo uso del algoritmo *One R attribute Eval*. Este evalúa el peso de cada atributo utilizando el algoritmo *One R.* Una vez obtenidos los valores de cada atributo se

aplica un *Ranker*, reduciendo el conjunto inicial. Por consiguiente, se consigue un subconjunto final que poseerá un nivel expresivo similar al original.

Esta operación se encuentra disponible dentro del *Submenu Attributes*, el cual se encuentra contenido dentro *Menú de Filtrado*. A continuación se muestra una instantánea de las opciones disponibles.

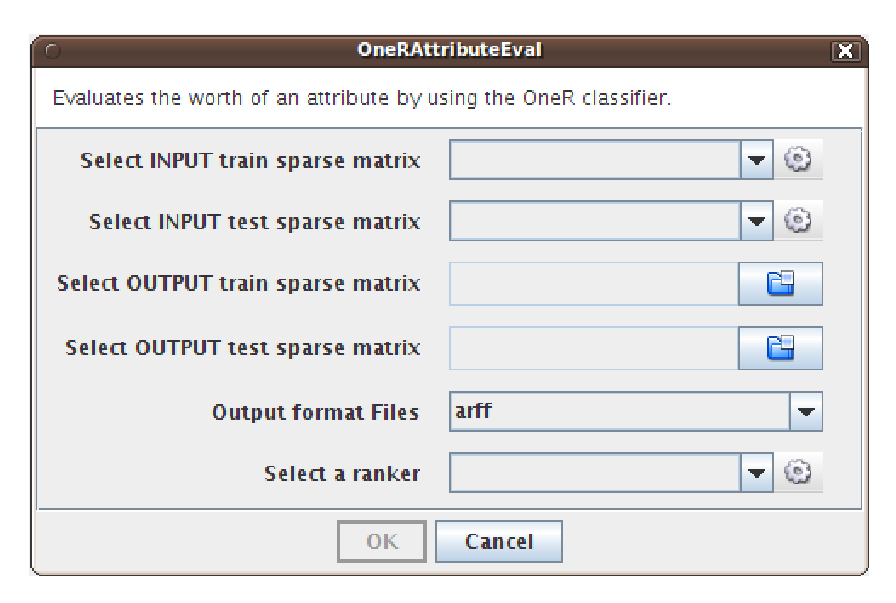


ILUSTRACIÓN 28 OPERACIÓN ONE R ATTRIBUTE EVAL

- **Select INPUT train sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de entrenamiento. Usando el desplegable, el usuario podrá seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select INPUT test sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de testeo. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select OUTPUT train sparse matrix:** Representa el fichero que se creará como nuevo corpus de entrenamiento una vez aplicado el filtro.

- **Select OUTPUT test sparse matrix:** Representa el fichero que se creará como nuevo corpus de testeo una vez aplicado el filtro.

- **Output format files:** El usuario puede seleccionar mediante el desplegable el tipo de fichero de salida: CSV o Arff. Por defecto el formato se ha preestablecido a *Arff.*

- **Ranker:** Dado que el filtro necesita establecer un rango de valores aceptados por el usuario, a través de esta opción, el usuario puede seleccionar uno creado con anterioridad o servirse del menú contextual para generar uno nuevo.

2.3.9 Operación Principal Components

La siguiente operación permite al usuario realizar un filtrado de atributos sobre un conjunto de datos haciendo uso del algoritmo *Principal Components*. Este realiza un análisis de

componentes principales generando un nuevo conjunto de datos más reducido y con pesos asociados. Una vez acabado el proceso se aplicará un *Ranker*, reduciendo el conjunto en función de los umbrales establecidos. Por consiguiente se consigue un subconjunto final que poseerá un nivel expresivo similar al original.

Esta operación se encuentra disponible dentro del *Submenu Attributes*, el cual se encuentra contenido dentro *Menú de Filtrado*. A continuación se muestra una instantánea de las opciones disponibles.

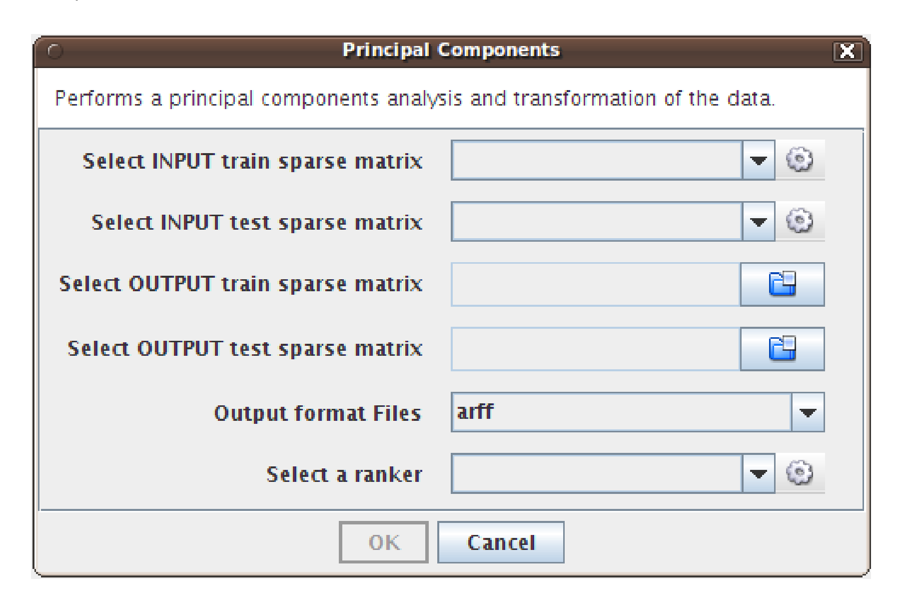


ILUSTRACIÓN 29 OPERACIÓN PRINCIPAL COMPONENTS

- **Select INPUT train sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de entrenamiento. Usando el desplegable, el usuario podrá seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select INPUT test sparse matrix:** Este parámetro se corresponde con el corpus que se utilizaría como conjunto de testeo. Usando el desplegable, el usuario puede seleccionar entre los corpus existentes en la aplicación o cargar uno nuevo usando el menú contextual.

- **Select OUTPUT train sparse matrix:** Representa el fichero que se creará como nuevo corpus de entrenamiento una vez aplicado el filtro.

- **Select OUTPUT test sparse matrix:** Representa el fichero que se creará como nuevo corpus de testeo una vez aplicado el filtro.

- **Output format files:** El usuario puede seleccionar mediante el desplegable el tipo de fichero de salida: CSV o Arff. Por defecto el formato se ha preestablecido a *Arff.*

- **Ranker:** Dado que el filtro necesita establecer un rango de valores aceptados por el usuario, a través de esta opción, el usuario puede seleccionar uno creado con anterioridad o servirse del menú contextual para generar uno nuevo.

2.3.10 CREAR UN RANKER

Esta operación permite al usuario crear una ordenación de atributos que sirva como umbral de valores en combinación con algunos tipos de filtros. Esta opción se encuentra disponible a través de los menús contextuales de los filtros, o también desde el *Submenu Attributes* perteneciente al grupo *Filtering.* A continuación se muestra una imagen de la interfaz de usuario asociada.

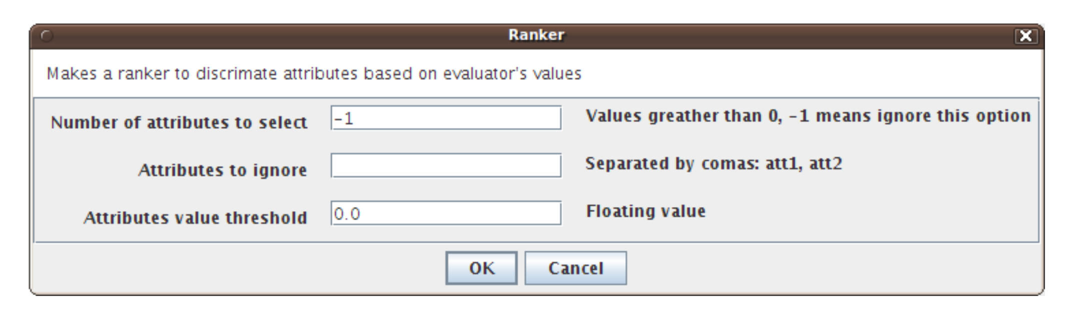


ILUSTRACIÓN 30 OPERACIÓN CREAR RANKER

Como se puede observar en la figura, esta posee varias opciones de configuración.

- **Number of attributes to select:** Permite al usuario establecer el número máximo de atributos a considerar en el conjunto final. Los valores están restringidos a enteros iguales o mayores que *0*, considerando de forma especial el valor *-1*, el cual permite ignorar la opción. En caso de introducir valores erróneos se notificará al usuario mediante una ventana de información. Por defecto, la opción se encuentra a -1.

- **Attributes to ignore:** Conjunto de atributos (separados por comas) que se ignorarán del ranking. Por tanto significa su directa exclusión del subconjunto final.

- **Attributes value threshold:** Opción más importante del *Ranker* pues establece el umbral de discriminación de valores del subconjunto final. Todos aquellos valores de atributos que no lo superen son discriminados. Permite valores flotantes sin restricción de valores numéricos. En caso de introducir valores no numéricos se notificará al usuario mediante una ventana de información.

3 VISORES

Como se comentó en los aspectos generales de AIBench, los visores cubren la parte relacionada con las interfaces de usuario. De esta manera se han creado visores para dar soporte a la modificación de los elementos creados o a los resultados de las operaciones llevadas a cabo.

En esta sección se ha considerado el separar los visores en dos tipos diferentes, de operaciones y de resultados, para su mejor compresión.

3.1 VISORES DE OPERACIONES

Los visores de operaciones permiten al usuario interactuar con objetos creados mediante el uso de operaciones, permitiendo entre otras cosas su modificación para usos posteriores. Dentro de estos Visores se engloban los de creación de modelos de clasificación y tipos de Kernel.

Estos visores poseen características homogéneas con el fin de mejorar la usabilidad:

- Todos los visores poseen paneles que agrupan la información en función de su tipo, a saber: opciones, ayuda y panel botones de acción.
- **Panel de opciones:** Recogen las opciones que se puede modificar del elemento, en algunos casos el visor puede poseer opciones de carácter explicito.
- **Panel de ayuda:** Este panel presenta la ayuda asociada al elemento y a todas sus opciones.
- **Botones de acción:** Mediante los botones de acción, **Update** y **Reset**, el usuario puede actualizar los valores del elemento a los actuales o volver a su estado original. Si el usuario decide actualizar los datos, se realiza un proceso de verificación de los mismos, en el cual si se produce algún error se notificará mediante una ventana informativa.

3.1.1 VISOR NAIVE BAYES

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un modelo *Naive Bayes*. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*; es decir, es necesario seleccionar un objeto del tipo *Naive Bayes* para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

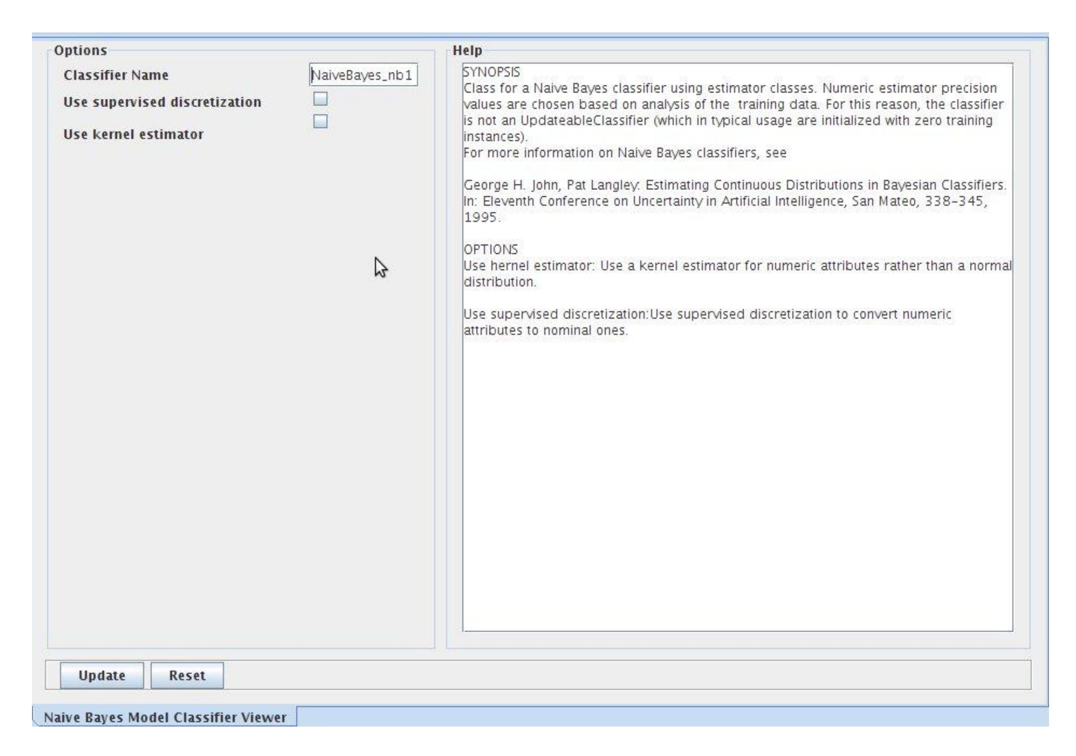


ILUSTRACIÓN 31 VISOR NAIVE BAYES

Como se puede observar en la imagen, el visor presenta las mismas opciones que se podían configurar mediante la operación *Crear modelo Naive Bayes*.

3.1.2 VISOR K NEAREST NEIGHBOR

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un modelo K Nearest Neighbor. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*; es decir, es necesario seleccionar un objeto del tipo K Nearest Neighbor para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

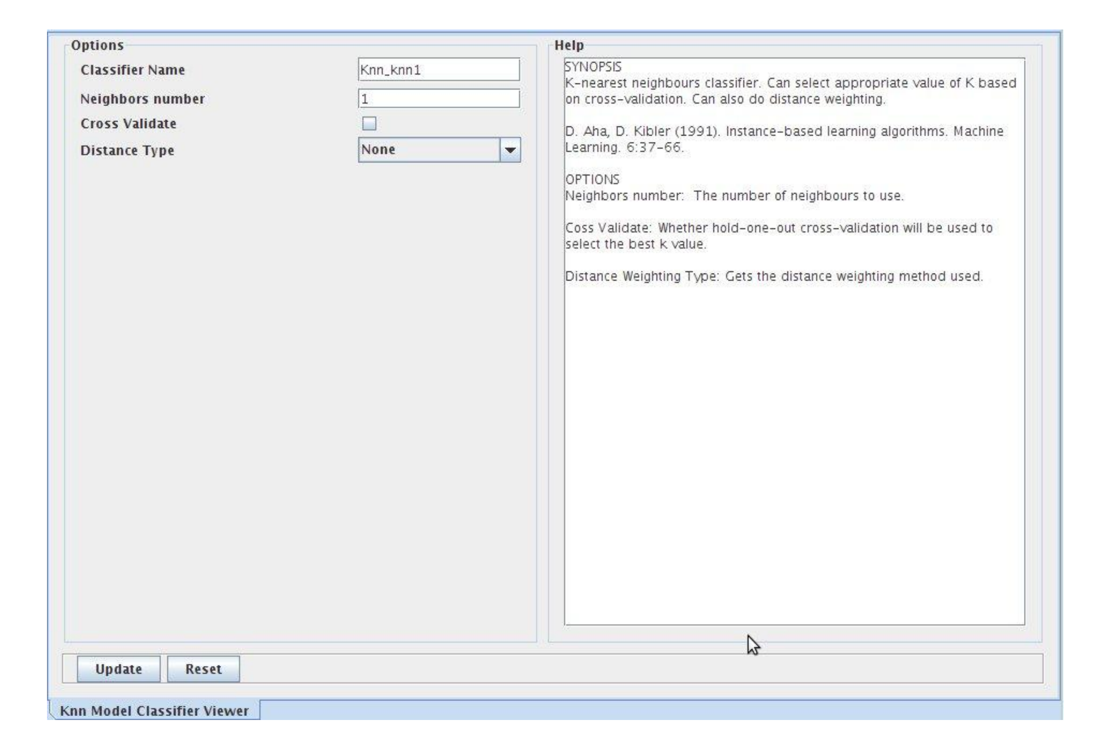


ILUSTRACIÓN 32 VISOR K NEAREST NEIGHBOR

Como se puede observar en la imagen, el visor presenta las mismas opciones que se podían configurar mediante la operación *Crear modelo* K Nearest Neighbor.

3.1.3 VISOR COST SUPPORT VECTOR MACHINE

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un modelo Cost Support Vector Machine. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*. Es decir, es necesario seleccionar un objeto del tipo Cost Support Vector Machine para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

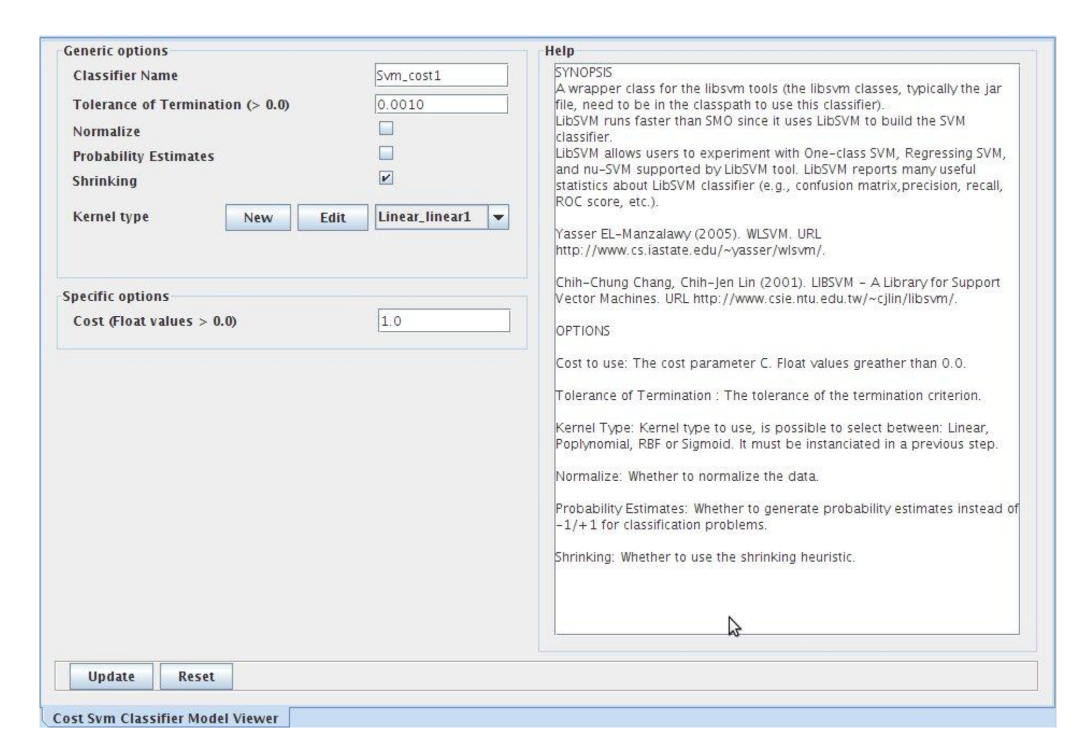


ILUSTRACIÓN 33 VISOR COST SUPPORT VECTOR MACHINE

Como se puede observar en la imagen, el visor presenta dos tipos de opciones, generales y especificas. Esto es debido a que BioClass soporta múltiples tipo de SVMs sobre la misma API. Aun así, estas no dejan de corresponderse con las que se podían configurar mediante la operación de creación.

En este caso particular, se han añadido dos botones **New** y **Edit**, asociados al parámetro Kernel Type:

- **Botón New:** Haciendo uso de este botón, el usuario puede crear nuevos Kernels, pues enlazan directamente con las operaciones disponibles a través de los menús.

- **Boton Edit:** Este botón lanza el visor asociado al tipo de Kernel seleccionado en el desplegable que se encuentra a su derecha.

2.2.1

3.1.4 VISOR NU SUPPORT VECTOR MACHINE

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un modelo Nu Support Vector Machine. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*. Es decir, es necesario seleccionar un objeto del tipo Nu Support Vector Machine para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

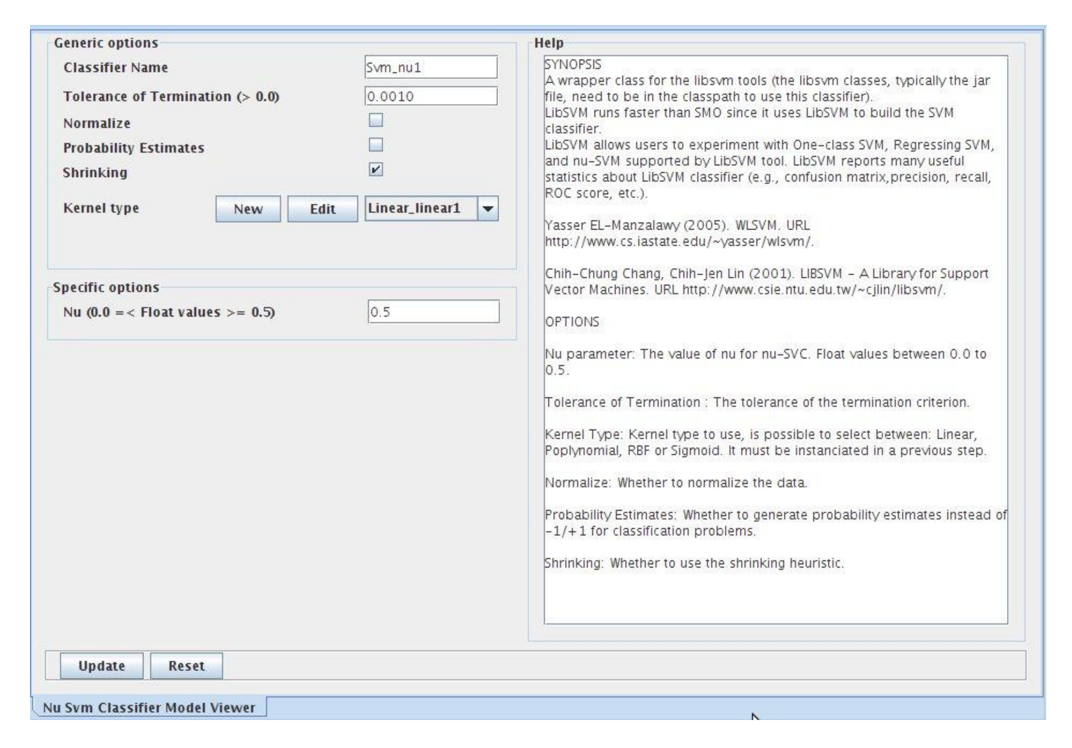


ILUSTRACIÓN 34 VISOR NU SUPPORT VECTOR MACHINE

Como se puede observar en la imagen, el visor presenta dos tipos de opciones, generales y especificas, esto es debido a que BioClass soporta múltiples tipo de SVMs sobre la misma API. Aun así, estas no dejan de corresponderse con las que se podían configurar mediante la operación de creación.

En este caso particular, se han añadido dos botones **New** y **Edit**, asociados al parámetro Kernel Type:

- **Botón New:** Haciendo uso de este botón el usuario puede crear nuevos *Kernels*, pues enlazan directamente con las operaciones disponibles a través de los menús.

- **Boton Edit:** Este botón lanza el visor asociado al tipo de *Kernel* seleccionado en el desplegable que se encuentra a su derecha.

2.2.1

3.1.5 VISOR KERNEL LINEAL

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un *Kernel Lineal*. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*. Es decir, es necesario seleccionar un objeto del tipo *Kernel Lineal* para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

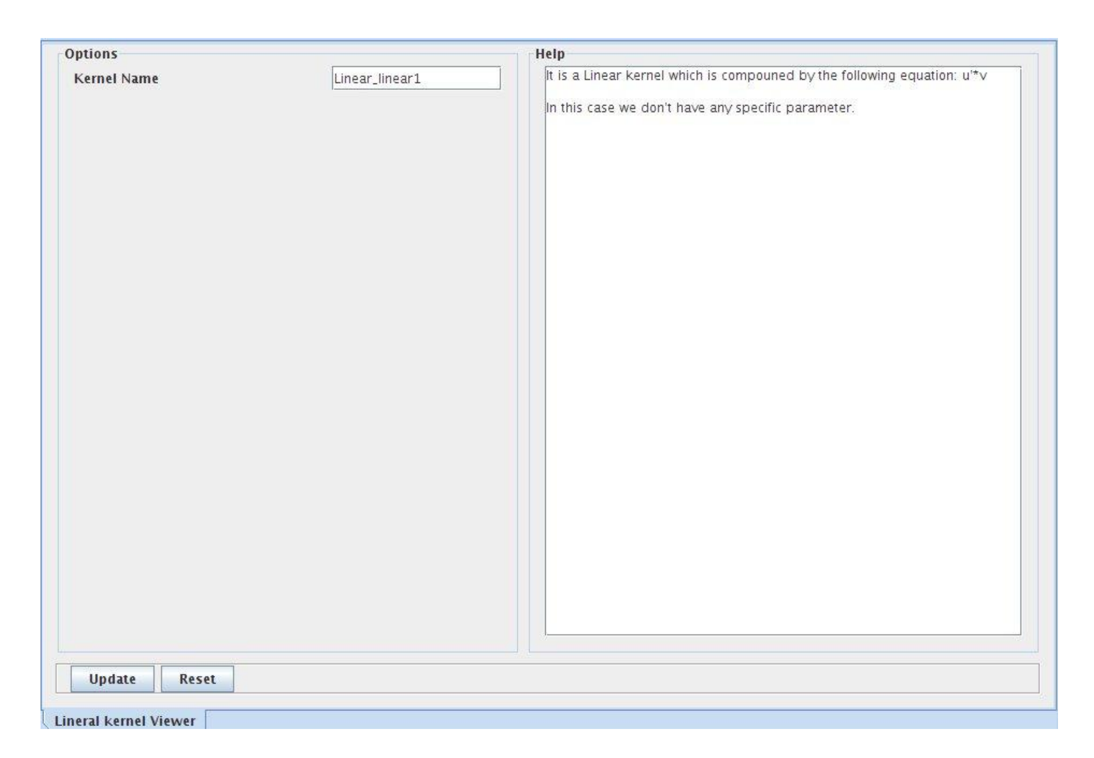


ILUSTRACIÓN 35 VISOR KERNEL LINEAL

Como se puede observar en la imagen, el visor presenta las mismas opciones que se podían configurar mediante la operación *Crear Kernel Lineal*.

3.1.6 VISOR KERNEL POLINOMIAL

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un Kernel Polinomial. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*, es decir, es necesario seleccionar un objeto del tipo Kernel Polinomial para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

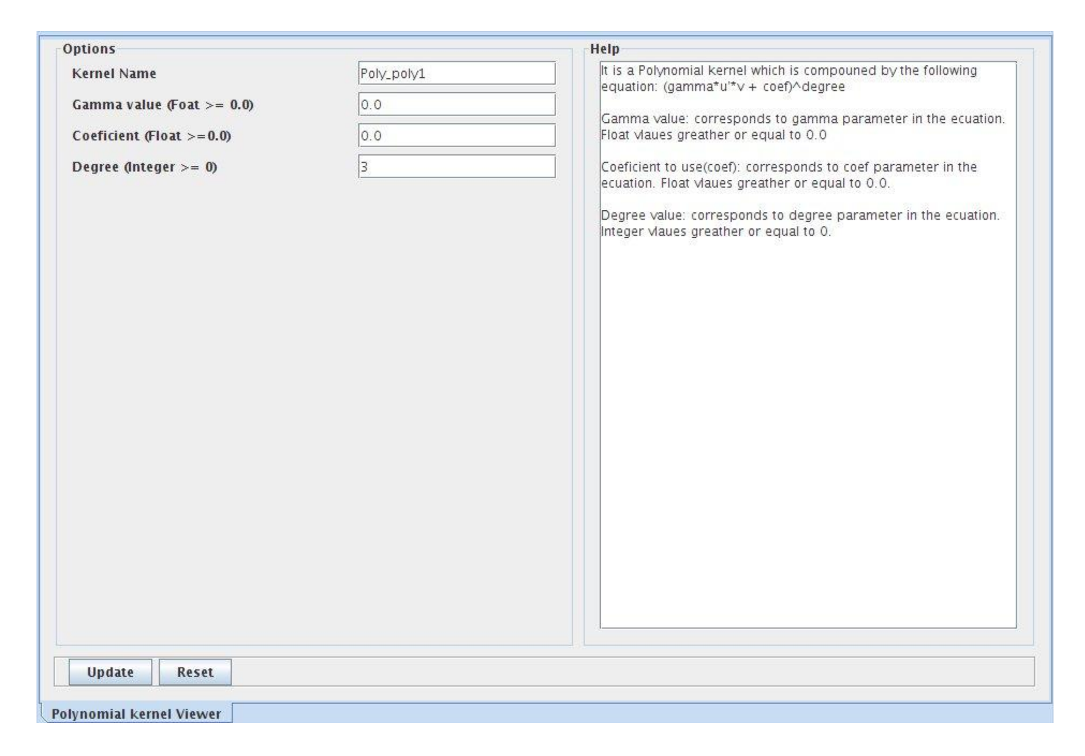


ILUSTRACIÓN 36 VISOR KERNEL POLINOMIAL

Como se puede observar en la imagen, el visor presenta las mismas opciones que se podían configurar mediante la operación *Crear Kernel Polinomial*.

3.1.7 Visor Kernel Radial

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un *Kernel Radial*. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*. Es decir, es necesario seleccionar un objeto del tipo *Kernel Radial* para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

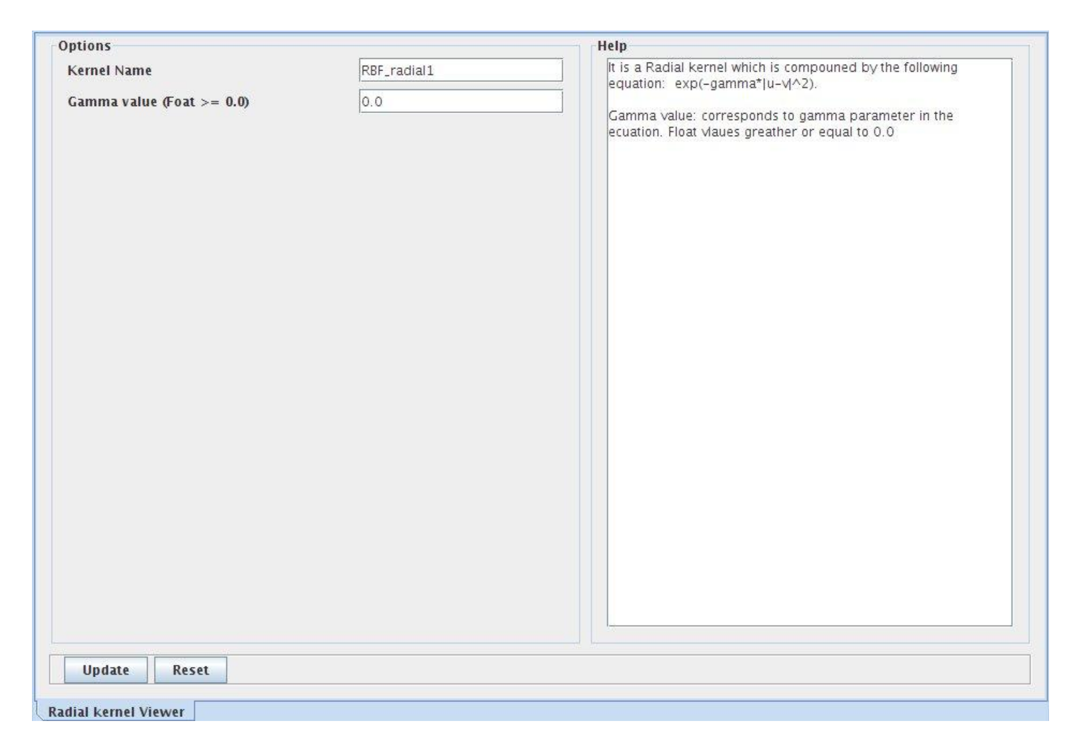


ILUSTRACIÓN 37 VISOR KERNEL RADIAL

Como se puede observar en la imagen, el visor presenta las mismas opciones que se podían configurar mediante la operación *Crear Kernel con base radial*.

3.1.8 VISOR KERNEL SIGMOIDAL

Utilizando este visor el usuario puede modificar y visualizar los parámetros asociados a un *Kernel Sigmoidal*. El proceso de visualización viene determinado por el proceso de interacción del usuario con el *Panel del Clipboard*. Es decir, es necesario seleccionar un objeto del tipo *Kernel Sigmoidal* para que se cargue el visor. A continuación se muestra una imagen de dicho visor.

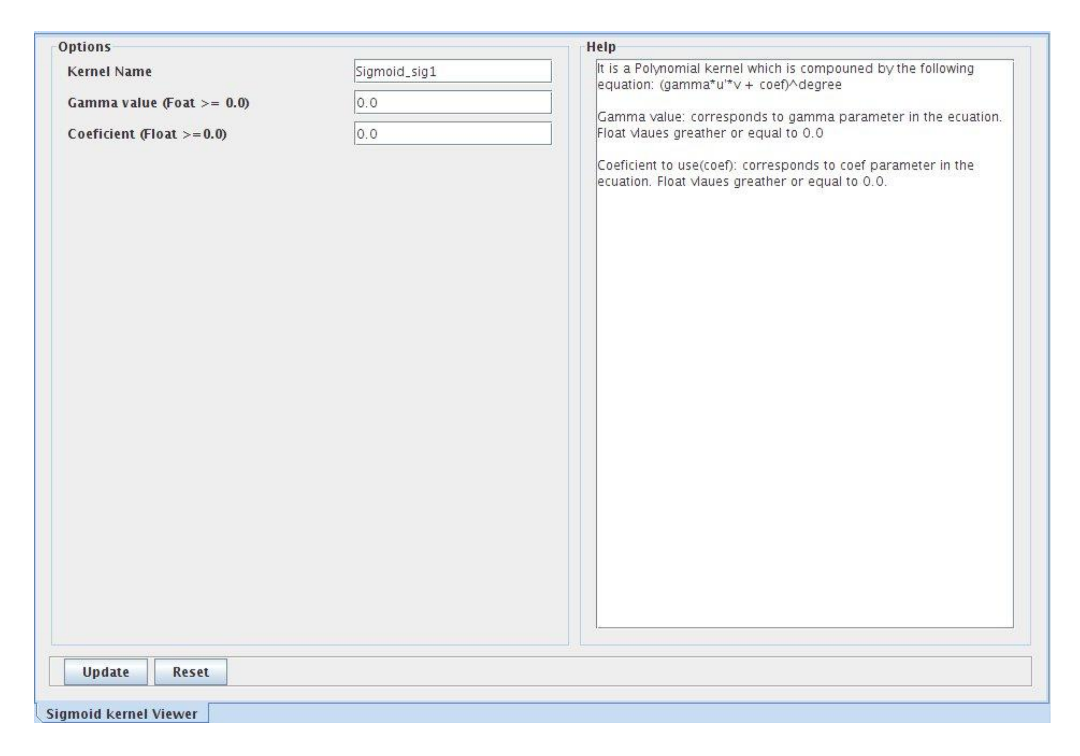


ILUSTRACIÓN 38 VISOR KERNEL SIGMOIDAL

Como se puede observar en la imagen, el visor presenta las mismas opciones que se podían configurar mediante la operación *Crear Kernel Sigmoidal*.

3.2 VISOR DE LA MATRIZ DE DISPERSIÓN

El visor de la matriz de dispersión, permite al usuario inspeccionar los datos contenidos en un corpus determinado. El método de lanzamiento es a través del *Panel del Clipboard*. Para ello es necesario seleccionar cualquier elemento de tipo matriz.

Previo a una explicación de la interfaz, se muestran a continuación unas vistas generales de la misma.

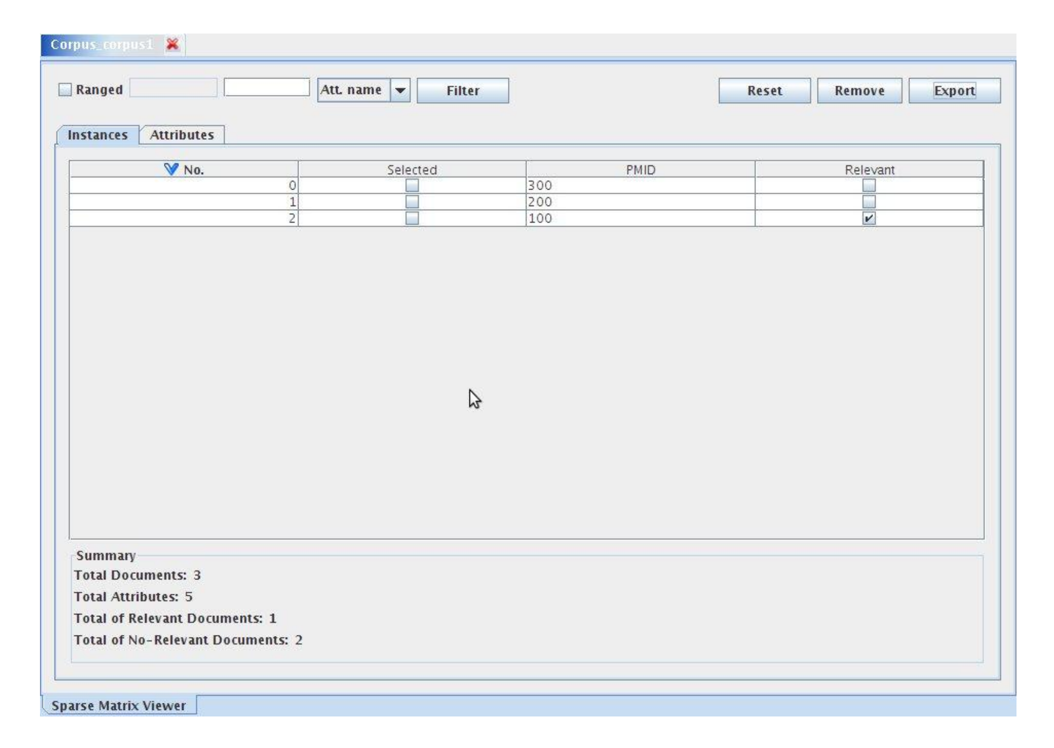


ILUSTRACIÓN 39 VISTA GENERAL VISOR MATRIZ DE DISPERSIÓN (INSTANCIAS)

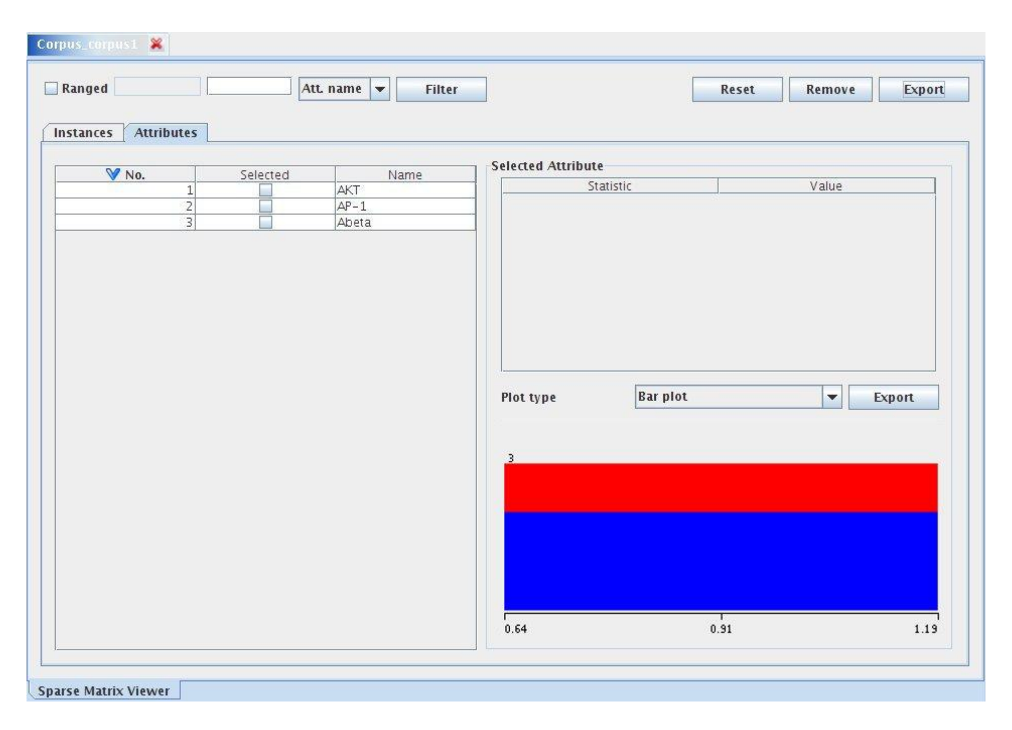


ILUSTRACIÓN 40 VISTA GENERAL VISOR MATRIZ DE DISPERSIÓN (ATRIBUTOS)

Como se puede ver en las ilustraciones, la interfaz cuenta con varios paneles de funcionamiento:

- **Barra de botones:** Esta barra permite al usuario manipular los datos contenidos dentro de la matriz. Desde aquí, se pueden realizar acciones de filtrado, eliminación, reinicio o exportación de los datos actuales.
- **Panel de datos:** El panel de datos representa el contenido de la matriz de dispersión, separando esta por los tipos de datos que gestiona (Atributos e Instancias).

3.2.1 BARRA DE BOTONES

Como se ha comentado previamente la barra de botones permite al usuario interactuar con los datos contenidos en la matriz. Veamos más detenidamente la función de cada uno de los botones.

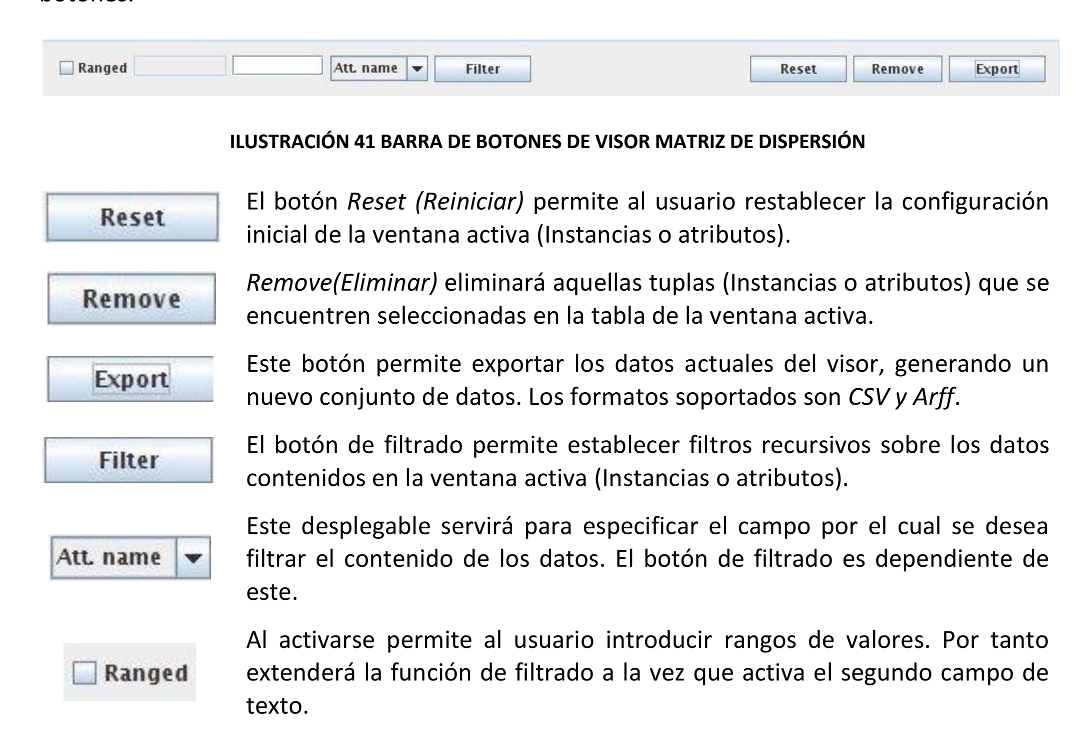


ILUSTRACIÓN 41 BARRA DE BOTONES DE VISOR MATRIZ DE DISPERSIÓN

El botón *Reset (Reiniciar)* permite al usuario restablecer la configuración inicial de la ventana activa (Instancias o atributos).

Remove(Eliminar) eliminará aquellas tuplas (Instancias o atributos) que se encuentren seleccionadas en la tabla de la ventana activa.

Este botón permite exportar los datos actuales del visor, generando un nuevo conjunto de datos. Los formatos soportados son *CSV y Arff*.

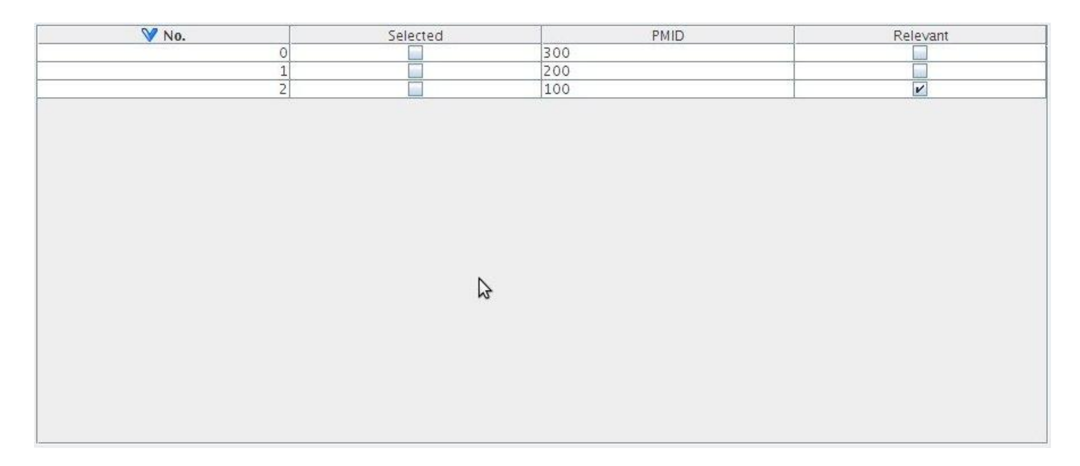
El botón de filtrado permite establecer filtros recursivos sobre los datos contenidos en la ventana activa (Instancias o atributos).

Este desplegable servirá para especificar el campo por el cual se desea filtrar el contenido de los datos. El botón de filtrado es dependiente de este.

Al activarse permite al usuario introducir rangos de valores. Por tanto extenderá la función de filtrado a la vez que activa el segundo campo de texto.

3.2.2 PANEL DE DATOS (INSTANCIAS)

El panel de instancias se encuentra dividido en dos secciones, una tabla que recoge las instancias existentes de la matriz de dispersión y otra que muestra un resumen estadístico. Veamos con detenimiento la primera, tal y como se muestra en la siguiente ilustración.



No.	Selected	PMID	Relevant
0	☐	300	☐
1	☐	200	☐
2	☐	100	☑

ILUSTRACIÓN 42 TABLA DE INSTANCIAS

La tabla dispone de varias características importantes, pues permite ordenar las tuplas por orden ascendente o descendente, como también seleccionar aquellas de deseen desechar.

- **No:** indica el número de instancia a mostrar según el orden de la matriz de dispersión.

- **Selected:** permitirá al usuario seleccionar aquellas instancias en las cuales este interesado en realizar alguna acción, como por ejemplo eliminarlas.

- **PMID:** En caso de haber cargado el corpus especificando el parámetro de identificación de documento, se añadirá a la tabla y al desplegable de filtrado pudiendo operar sobre ello. El nombre del identificador variará en función del que se haya especificado.

- **Relevant:** Indicará la relevancia de esa instancia en relación al corpus.

La segunda sección del panel, como se comentó, muestra un pequeño resumen estadístico de la matriz de dispersión, a saber: **Número de documentos, atributos, documentos relevantes y no relevantes.**

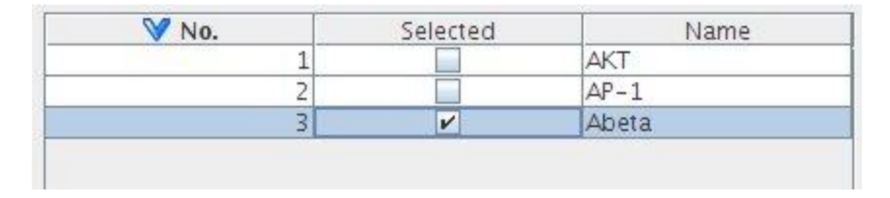
Summary
Total Documents: 3
Total Attributes: 5
Total of Relevant Documents: 1
Total of No-Relevant Documents: 2

ILUSTRACIÓN 43 SUMARIO DE MATRIZ

3.2.3 Panel de datos (Atributos)

La segunda parte del parte del panel de datos esta reservado a los atributos relevantes a cada documento. Al igual que el panel de Instancias, este también posee varias secciones bien diferenciadas: Ilustración 44 tabla de atributos, Ilustración 45 tabla de estadisticos por atributo e Ilustración 46 gráficas de atributo.

Veamos detenidamente cada una de esas ilustraciones junto con una explicación de cada una de las opciones que ofrece.

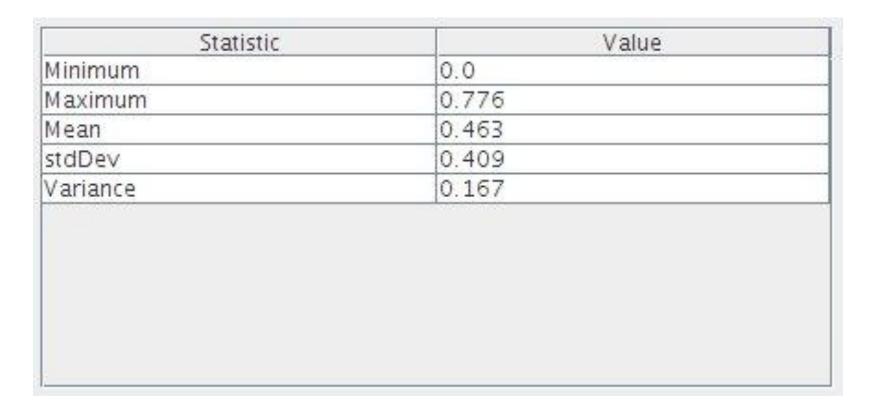


No.	Selected	Name
1	☐	AKT
2	☐	AP-1
3	✔	Abeta

ILUSTRACIÓN 44 TABLA DE ATRIBUTOS

- **No:** indica el número de instancia a mostrar según el orden de la matriz de dispersión.
- **Selected:** permite al usuario seleccionar aquellos atributos, en los cuales este interesado en realizar alguna acción, como por ejemplo eliminarlos.
- **Name:** Representa el nombre del atributo en cuestión.

En este caso concreto, la tabla de atributos se encuentra íntimamente ligada a la tabla de estadísticos, pues el usuario puede seleccionar un atributo en la primera para que se muestren los estadísticos de este en la segunda.



Statistic	Value
Minimum	0.0
Maximum	0.776
Mean	0.463
stdDev	0.409
Variance	0.167

ILUSTRACIÓN 45 TABLA DE ESTADISTICOS POR ATRIBUTO

Como se puede ver en la ilustración de estadísticos, se han recogido aquellos más conocidos y relevantes para este tipo de estudio: **Mínimo, Máximo, Media, Desviación típica y Varianza**.

Por último, el panel de atributos también dispone de una sección destinada a la generación de gráficas. Esta permite al usuario observar los experimentos desde una perspectiva más visual e intuitiva. La siguiente imagen muestra como se vería dicha sección.

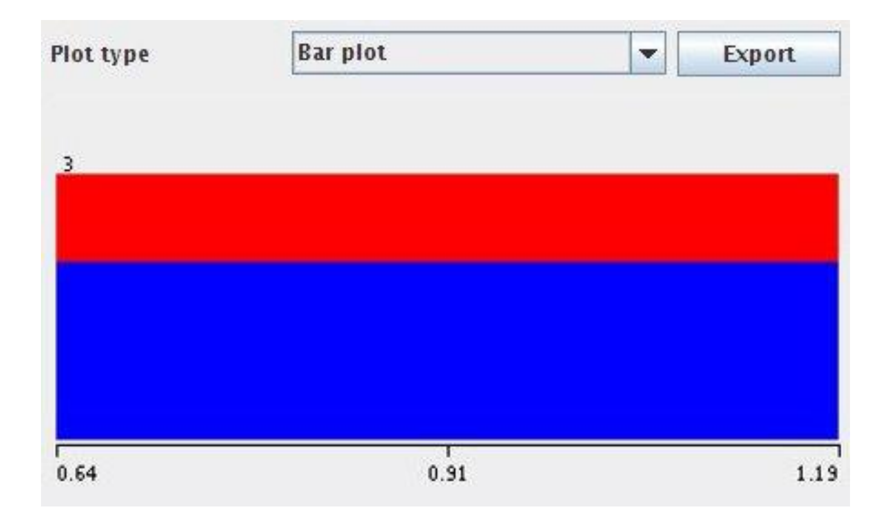


ILUSTRACIÓN 46 GRÁFICAS DE ATRIBUTO



El desplegable permitirá seleccionar el tipo de grafico que se desee mostrar. Actualmente están soportados los gráficos de barras y dispersión.



Este botón permite exportar el gráfico actual a un fichero de imagen. Los formatos soportados son *PNG y JPEG*.

3.3 VISOR DE RESULTADOS DE CLASIFICACIÓN

Se ha creado un visor asociado a la salida de un clasificador. Del mismo modo que se han lanzado otros visores, este se ejecuta de la misma manera. Bastará con seleccionar un elemento de tipo resultado sobre el *Panel del Clipboard*.

La siguiente ilustración muestra una visión general del panel de resultados.

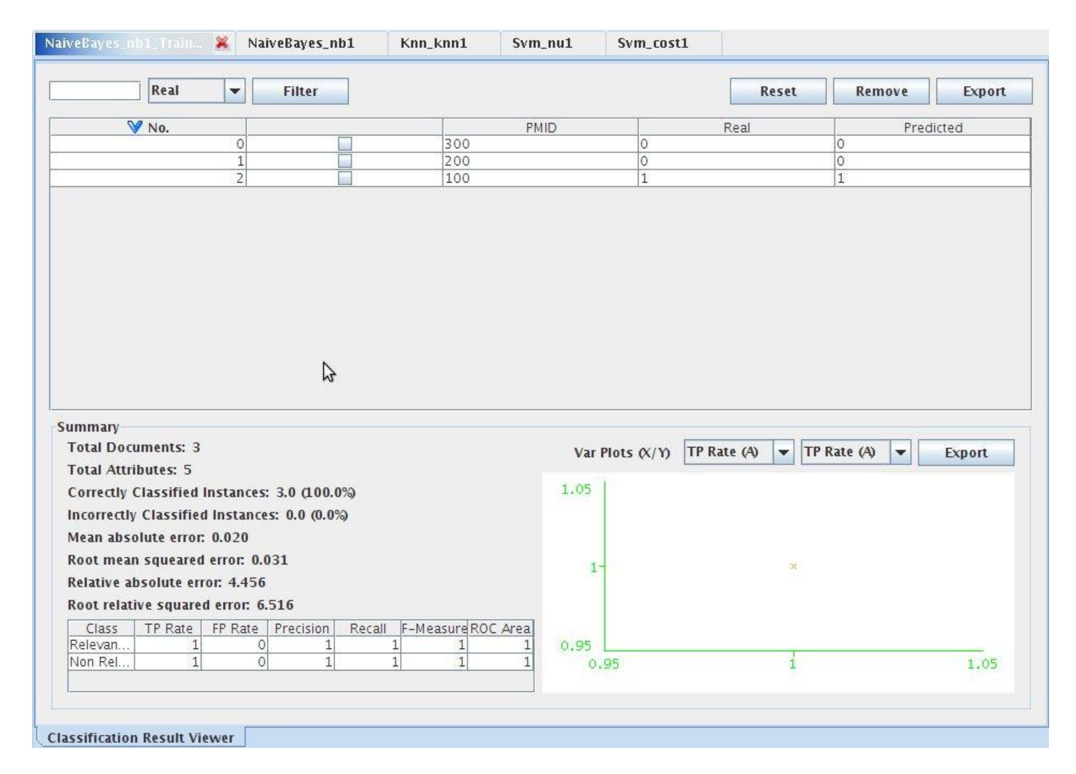


ILUSTRACIÓN 47 VISTA GENERAL PANEL DE RESULTADOS

Como se puede ver en la ilustración, la interfaz cuenta con varias secciones agrupadas por funcionamiento:

- **Barra de botones:** Esta barra permite al usuario manipular los datos obtenidos como resultados del proceso de clasificación. Desde aquí, se podrán realizar acciones de filtrado, eliminación, reinicio o exportación de los datos actuales.
- **Tabla de predicciones:** La tabla de predicciones recoge los resultados comparativos del proceso de clasificación de cada instancia.
- **Sumario:** Esta sección recoge todos los datos estadísticos asociados con el proceso de clasificación, permitiendo también generar y exportar gráficos de dispersión.

3.3.1 BARRA DE BOTONES

Como se ha comentado previamente, la barra de botones permite al usuario interactuar con los datos contenidos en la matriz.

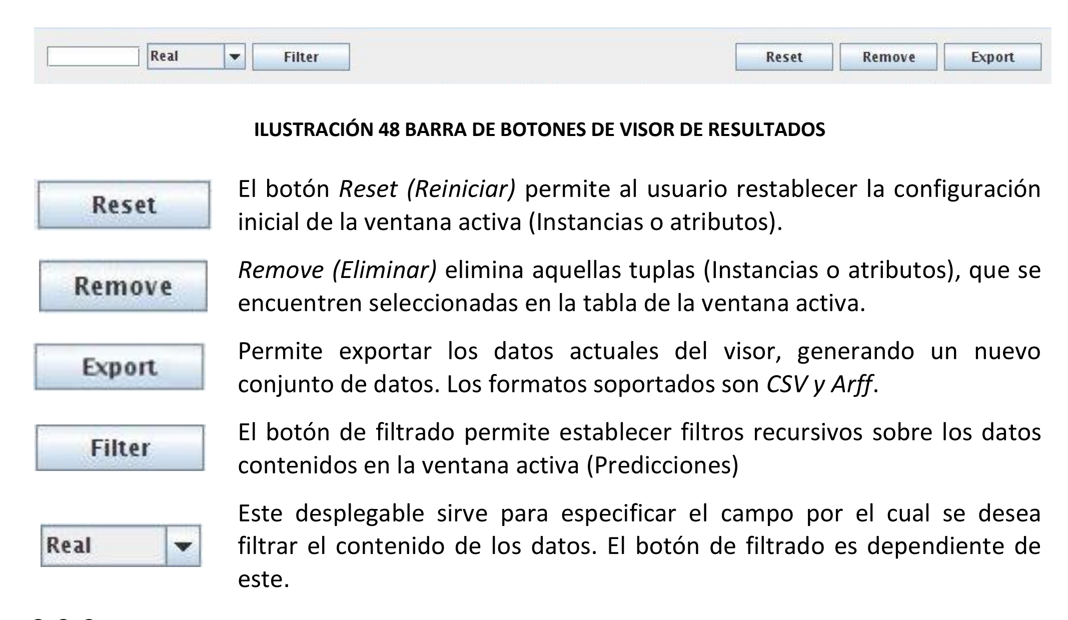


ILUSTRACIÓN 48 BARRA DE BOTONES DE VISOR DE RESULTADOS

Reset: El botón *Reset (Reiniciar)* permite al usuario restablecer la configuración inicial de la ventana activa (Instancias o atributos).

Remove: *Remove (Eliminar)* elimina aquellas tuplas (Instancias o atributos), que se encuentren seleccionadas en la tabla de la ventana activa.

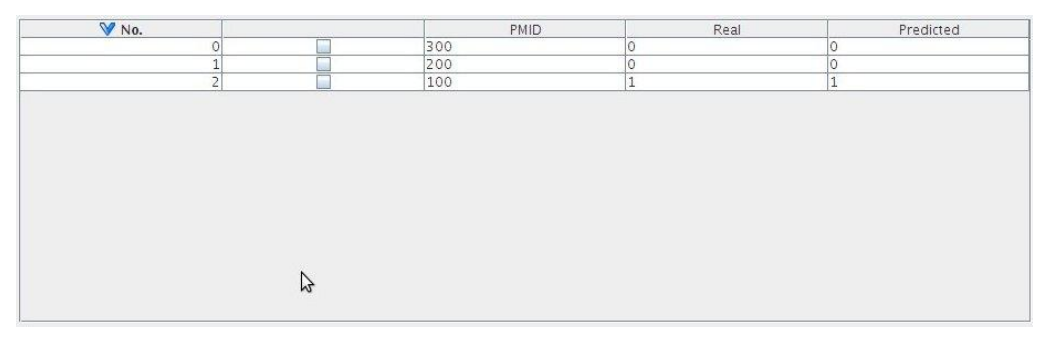
Export: Permite exportar los datos actuales del visor, generando un nuevo conjunto de datos. Los formatos soportados son *CSV y Arff*.

Filter: El botón de filtrado permite establecer filtros recursivos sobre los datos contenidos en la ventana activa (Predicciones)

Real: Este desplegable sirve para especificar el campo por el cual se desea filtrar el contenido de los datos. El botón de filtrado es dependiente de este.

3.3.2 TABLA DE PREDICCIONES

Como se comentó, la tabla de predicciones muestra los resultados obtenidos de la clasificación a nivel de instancias, en contraposición de los resultados reales (en caso de que existieran). La siguiente ilustración muestra lo comentado.



No.		PMID	Real	Predicted
0	☐	300	0	0
1	☐	200	0	0
2	☐	100	1	1

ILUSTRACIÓN 49 TABLA DE PREDICCIONES

La tabla dispone de varias características importantes, pues permite ordenar las tuplas por orden ascendente o descendente, como también seleccionar aquellas de deseen desechar.

- **No:** indica el número de instancia a mostrar según el orden de la matriz de dispersión.
- **Selected:** permite al usuario seleccionar aquellas instancias en las cuales este interesado en destacar por algún motivo, como por ejemplo eliminarlas.
- **PMID:** En caso de haber cargado el corpus especificando el parámetro de identificación de documento, se añadirá a la tabla y al desplegable de filtrado pudiendo operar sobre ello. El nombre del identificador variará en función del que se haya especificado.

- **Real:** Indica el valor real de relevancia que debería tener la instancia. Este campo servirá como contraste del proceso de clasificación.

- **Predicted:** Indica el valor de relevancia predicho por el clasificador.

3.3.3 SUMARIO

La sección de sumario recoge el análisis estadístico realizado una vez terminado el proceso de clasificación. Este se encuentra separado a su vez en dos partes bien diferenciadas, estadísticos a la izquierda, generación de gráficos a la derecha. Veamos una instantánea de la sección.

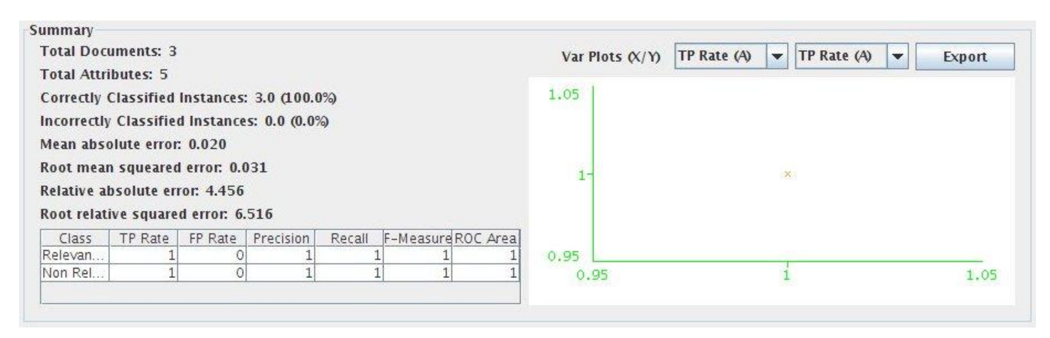
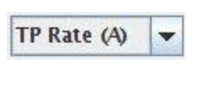


ILUSTRACIÓN 50 SUMARIO DE RESULTADOS Y GRÁFICAS

Según la ilustración, la parte izquierda mostraría los siguientes estadisticos:

- **Total documents:** indica el número total de documentos contenidos en el corpus.
- **Total Attributes:** indica el número total de atributos que contiene el corpus.
- **Correctly Classified Instances:** número y porcentaje (entre paréntesis) de instancias correctamente clasificadas.
- **Incorrectly Classified Instances:** número y porcentaje (entre paréntesis) de instancias incorrectamente clasificadas.
- **Mean absolute error:** indica el error medio absoluto.
- **Root mean squared error:** indica el error cuadrático medio.
- **Relative absolute error:** indica el error absoluto relativo.
- **Root relative squeared error:** indica el error cuadrático relativo.
- **Tabla de estadísticos por clase:** Indica los estadísticos más utilizados en los procesos de clasificación organizados por clase. **TP Rate(Porcentaje de verdaderos positivos), FP Rate(Porcentaje de verdaderos negativos), Precision, Recall, F-Measure y Area ROC.**

La parte derecha en cambio permite al usuario combinar los distintos estadísticos de cada clase para generar un grafico asociado a los datos, el cual si se desea se puede exportar. La opciones gráficas por tanto serían las siguientes:



Combinando los dos desplegables del mismo tipo el usuario puede establecer los datos que utilizan en los ejes de coordenadas. Mencionar que los estadísticos se encuentran separados por clase haciendo uso del siguiente nemotécnico: *Estadístico (CLASE).*



Este botón permite exportar el gráfico actual a un fichero de imagen. Los formatos soportados son *PNG y JPEG*.

4 ESCENARIOS

En esta sección se han recogido escenarios de uso de la aplicación. Se ha dividido en dos subsecciones las cuales tratar situaciones específicas acerca de la utilización de BioClass. La finalidad de esta pequeña guía,es servir como introducción al uso de la aplicación, pudiendo el usuario servirse de los ejemplos propuestos y crear los suyos propios.

4.1 ESCENARIO DE CLASIFICACIÓN

La Ilustración 51 muestra un escenario de uso de la herramienta BioClass en la cual se lleva a cabo un proceso completo de clasificación. A continuación se detallan cada uno de los pasos que se han llevado a cabo.

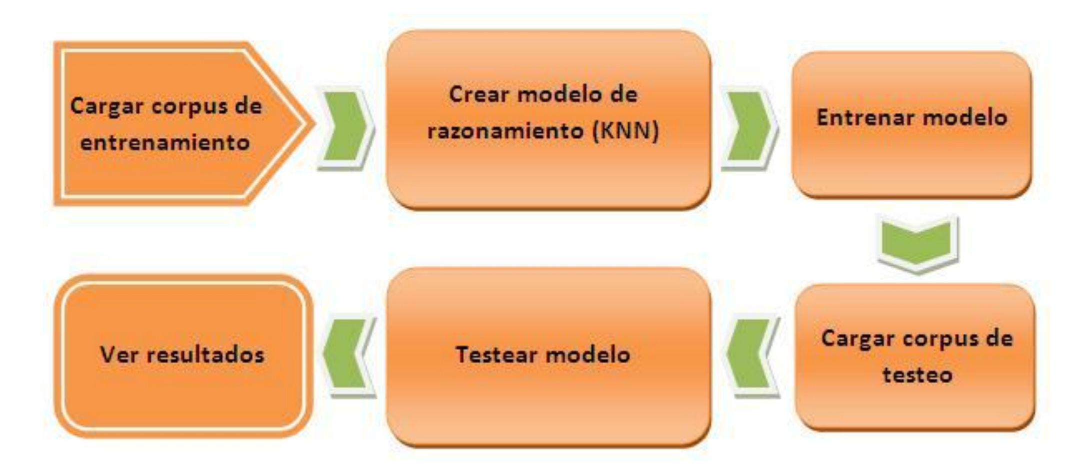


ILUSTRACIÓN 51 ESCENARIO DE CLASIFICACIÓN

1) **Cargar corpus de entrenamiento:** A través de la operación *Load corpus from file* del menú *Corpus* se ha cargado el conjuntos de datos de entrenamiento.
2) **Crear modelo de razonamiento (KNN):** El modelo de clasificación que ha usado esta basado en algoritmo de proximidad por vecindad KNN, para ello se ha utilizado la operación *K Nearest Neighbor* del menú *Classification.*
3) **Entrenar modelo:** Para entrenar el modelo se ha utilizado la operación *Train model* del menú *Classification* haciendo uso del modelo creado.
4) **Cargar corpus de testeo:** al igual que en el 1º paso, se ha utilizado la misma operación para cargar conjunto de datos a testear.
5) **Testear modelo:** Para testear el modelo ya entrenado sobre un conjunto de datos a testear, se ha utilizado la operación *Test model* del menú *Classification*.
6) **Ver resultados:** Una vez se ha terminado el proceso de clasificación, se pueden utilizar los visores de resultados para analizar los datos obtenidos. Para ello se debe hacer click sobre los objetos de tipo resultado en el *Panel del Clipboard*

4.2 ESCENARIO DE FILTRADO

La Ilustración 52 muestra un escenario de uso de la herramienta BioClass en la cual se lleva a cabo un proceso completo de filtrado de atributos. A continuación se detallan cada uno de los pasos que se han llevado a cabo.

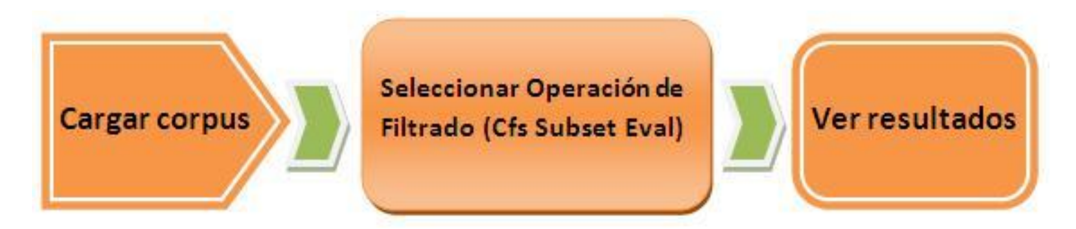


ILUSTRACIÓN 52 ESCEANARIO DE FILTRADO

1) **Cargar corpus de entrenamiento:** A través de la operación *Load corpus from file* del menú *Corpus* se ha cargado el conjuntos de datos de sobre el cual se desea aplicar un algoritmo de filtrado de atributos.
2) **Seleccionar operación de filtrado (Cfs Subset Eval):** El algoritmo de filtrado de atributos que se ha utilizado durante el proceso ha sido *Cfs Subset Eval*. A el se accede a través de la operación *Cfs Subset Eval,* del menú *Filtering*.
3) **Ver resultados:** Una vez terminado el proceso de clasificación, se pueden utilizar los visores de resultados para analizar los datos obtenidos. En este caso se usarán aquellos asociados a las matrices de dispersión, ya que el resultado de la operación produce nuevas matrices filtradas.